GBLUP and G matrices
GBLUP from SNP-BLUP

• We have defined breeding values as sum of SNP effects:
  \[ \mathbf{u} = \mathbf{Za} \]

• To refer breeding values to an average value of 0, we adopt the “centered” coding for genotypes described before

<table>
<thead>
<tr>
<th>Genotype</th>
<th>101 Coding</th>
<th>012 Coding</th>
<th>Centered coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>aa</td>
<td>(-a_i)</td>
<td>0</td>
<td>(-2p_i a_i)</td>
</tr>
<tr>
<td>Aa</td>
<td>0</td>
<td>(a_i)</td>
<td>((1 - 2p_i)a_i)</td>
</tr>
<tr>
<td>AA</td>
<td>(a_i)</td>
<td>2(a_i)</td>
<td>((2 - 2p_i)a_i)</td>
</tr>
</tbody>
</table>
GBLUP from SNP-BLUP

• We have defined breeding values as sum of SNP effects:
  \[ u = Za \]

• Because \( Var(\alpha) = I\sigma_a^2 \), then
  \[ Var(u) = Z(I\sigma_a^2)Z' = ZZ'\sigma_a^2 \]

• But before, we found out that \( \sigma_a^2 = \frac{\sigma_u^2}{2\sum p_i q_i} \). Substituting:
  \[ Var(u) = \frac{ZZ'}{2\sum p_i q_i} \sigma_u^2 \]

• Finally, we factorize \( \sigma_u^2 \)
VanRaden’s “first $G$”

$$G = \frac{(M-2P)(M-2P)'}{2\sum p_iq_i} = \frac{Z'Z}{2\sum p_iq_i}$$

Shifted to refer to the average of a population with allele frequencies $p$

Scaled to refer to the genetic variance of a population with allele frequencies $p$
GBLUP

\[
y = Xb + Wu + e
\]

\[
\begin{pmatrix}
X'R^{-1}X & X'R^{-1}W \\
W'R^{-1}X & W'R^{-1}W + G^{-1}\sigma_u^{-2}
\end{pmatrix}
\begin{pmatrix}
\hat{b} \\
\hat{u}
\end{pmatrix}
= \begin{pmatrix}
X'R^{-1}y \\
W'R^{-1}y
\end{pmatrix}
\]
GBLUP

\[
\begin{pmatrix}
X'R^{-1}X & X'R^{-1}W \\
W'R^{-1}X & W'R^{-1}W + G^{-1}\sigma_u^{-2}
\end{pmatrix}
\begin{pmatrix}
\hat{b} \\
\hat{u}
\end{pmatrix}
=
\begin{pmatrix}
X'R^{-1}y \\
W'R^{-1}y
\end{pmatrix}
\]

• We obtain animal, not SNP, solutions
• Immediate application to maternal effects model, random regression, competition effect models, multiple trait, etc.
• All genotyped individuals can be included, either with phenotype or not..
• Regular software (blupf90, asreml, wombat...) works
• Therefore, GREML and G-Gibbs are simple extensions.
Multiple trait GBLUP

\[
\begin{pmatrix}
X' R^{-1} X & X' R^{-1} W \\
W' R^{-1} X & W' R^{-1} W + G_0^{-1} \otimes G_0^{-1}
\end{pmatrix}
\begin{pmatrix}
\hat{b} \\
\hat{u}
\end{pmatrix}
= \begin{pmatrix}
X' R^{-1} y \\
W' R^{-1} y
\end{pmatrix}
\]

$G_0$ is the matrix of genetic covariance across traits usually $R = I \otimes R_0$, where $R_0$ is residual covariances.
Reliabilities

Nominal reliabilities (NOT cross-validation reliabilities) can be obtained from the Mixed Model equations, as:

\[ Rel_i = 1 - \frac{C^{ii}}{G_{ii}\sigma_u^2} \]

where \( C^{ii} \) is the \( i, i \) element of the inverse of the mixed model equations.
GBLUP == SNPBLUP

• Both give the same solutions
• We can jump from GBLUP to SNP-BLUP

\[ \hat{u} = Z\hat{\alpha} \]

\[ \hat{\alpha} = \frac{1}{2\sum p_i q_i} Z' G^{-1} \hat{u} \]
GREML, G-Gibbs...

Use of $G$ to estimate variance components...

It can be done with remlf90, gibbs*f90, AsReml, TM...

The result will refer to an ideal population with whatever allelic frequencies we introduced in the denominator of

$$G = \frac{z'z}{2\sum p_i q_i}.$$
But what are genomic (additive) relationships?
Kinship

**kin | kɪn | noun [treated as pl.]**

one's family and relations: *many elderly people have no kin to turn to for assistance.*

**ORIGIN**

Old English *cynn*, of Germanic origin; related to Dutch *kunne*, from an Indo-European root meaning ‘*give birth to*’, shared by Greek *genos* and Latin *genus* ‘*race*’. It obviously comes from Latin “parentes”
So what is kinship?

• Socially it has a “pedigree” interpretation
  • e.g. ”all royal families are related”

• However pedigrees “go back forever”

• We need a more rigorous definition
True relationships

• Two individuals are genetically identical (for a trait) if they carry the same genotype at the causal QTLs or genes
  • This is a *biological fact*

• The genetics of one locus for two diploid individuals can be described using Gillois’ identity coefficients
Relationships

• Relationships were conceived as standardized covariances (Fisher, Wright)
  • $\text{Cov}(u_i, u_j) = R_{ij}\sigma^2_u$
  • $R_{ij}$ “some” relationship
  • $\sigma^2_u$ genetic variance

• Genetic relationships are due to shared (Identical By State) alleles at causal genes
  • if I share the blood group 00 with somebody I am “like” his twin
  • These genes are unknown (and many will likely remain so)
  • Use proxies

• Pedigree relationships
• Marker relationships
Figure 2. The pedigree of the Jicaque Indians Julio and Mencha.
Pedigree relationships: A

- Systematic “tabular” rules to compute any $A_{ij}$ (Emik & Terrill 1947)
- The whole array of $A_{ij}$ is disposed in a matrix $A$.
- $A^{-1}$ is very sparse and easy to create and manipulate (Henderson 1976)
  - Extraordinary development of whole-pedigree methods in livestock genetics
  - E.g. computing inbreeding for 15 generations including $10^6$ sheep takes minutes
Early use of markers used them to infer $\mathbf{A}$

- In conservation genetics, molecular markers have often been used to estimate pedigree relationships

- Gather markers, then reconstruct pedigrees, then construct $\mathbf{A}$
  - Either estimates of $A_{xy}$, or estimates of « the most likely relation » (son-daughter, cousins, whatever)
  - Li and Horvitz 1953, Cockerham 1969, Ritland 1996, Caballero & Toro 2002, and many others

- With abundant marker data we can do better than this
Realized relationships

• Identical By Descent Relationships based on pedigree are average relationships which assume infinite loci.

• « Real » IBD relationships $R$ are a bit different due to finite genome size (Hill and Weir, 2010)

• Therefore $A$ is the expectation of realized relationships $R$

• SNPs more informative than $A$.
  • Two full sibs might have a correlation of 0.4 or 0.6

• You need many markers to get these « fine relationships »
Traditional Pedigree

Animal

- Sire
  - Sire of Sire
  - Dam of Sire
- Dam
  - Sire of Dam
  - Dam of Dam
Genomic Pedigree
Haplotype Pedigree

ctgtctagatcg
atgtcgcgcagtcagt

ctgtagcgatcg
agatctagatcg

ctgtagcttagg
agggcgcgcagt
cgatctagatcg
cggtagatcagt
agagatcgcagt
ctatcgctcagg

atagatcgatcg
ctgtagcttagg
agggcgcgcagt
cgatctagatcg
cggtagatcagt
agagatcgcagt
ctatcgctcagg
Genotype Pedigree

Count number of second allele

0 = homozygous for first allele (alphabetically)
1 = heterozygous
2 = homozygous for second allele (alphabetically)
Comparison of expected and observed variances – relationship/sharing

4401 full sib pairs
400-800 markers
Expected
Mean 0.5
SD 0.039

Observed
Mean 0.498
SD 0.036
Range 0.37 - 0.63

Source: Visscher et al.

Slide from WG Hill
Covariance of gene content

• Consider gene content coding \{AA, Aa, aa\} as 
  \[ m = \{01,2\} \]

• Cockerham, 1969:
  • For two individuals, the covariance of their gene contents is 
  \[ Cov(m_i, m_j) = R_{ij} 2pq \]
  • In other words, two related individuals will show similar genotypes at the markers

• Backsolve \( \hat{R}_{ij} = Cov(m_i, m_j)/2pq \).

• If we have centered \( z = m - 2p \) then \( \hat{R}_{ij} = \frac{z_iz_j}{2pq} \)

• Extended to many loci \( \hat{R}_{ij} = \frac{\text{mean}(z_iz_j')}{\text{mean}(2\sum p_kq_k)} = \frac{z_iz_j'}{2\sum p_kq_k} \)
VanRaden’s “first $G$”

If base allelic frequencies are used, $G$ is an unbiased estimator of IBD realized relationships.

$$ G = \frac{(M-2P)(M-2P)'}{2\sum p_iq_i} = \frac{Z'Z}{2\sum p_iq_i} $$

Shifted to refer to the average of a population with allele frequencies $p$

Scaled to refer to the genetic variance of a population with allele frequencies $p$
Some properties of $G$

- If $p$ are computed from the sample
- In HWE & Linkage Equilibrium
  - Average of $\text{Diag}(G) = 1$
  - Average $(G) = 0$
- With average inbreeding $F$
  - Average of $\text{Diag}(G) = 1 + F$

$$G = \frac{(M-2P)(M-2P)'}{2\sum p_iq_i}$$

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>freq</td>
<td>$q^2 + pqF$</td>
<td>$2pq(1-F)$</td>
<td>$p^2 + pqF$</td>
</tr>
</tbody>
</table>
Some intriguing properties of $G$

• If $p$ are computed from the data
  • This implies that $E(\text{Breeding Values})=0$

• Positive and negative inbreeding
  • Some individuals are more heterozygous than the average of the population (OK, no biological problem)

• Positive and negative genomic relationships
  • This implies that individuals $i$ and $j$ are more distinct than an average pair of individuals in the data
  • Fixing negative estimates of relationships to 0 is wrong praxis
Not positive definite

- Strandén & Christensen (2011) showed that if $p$’s are averages across the sample then $G$ is not positive definite (has no inverse).

- We could use BLUP equations with non-inverted $G$ (Henderson, 1984) => see exercises.

- Instead, we use $G = 0.99 \frac{(M-2P)(M-2P)'}{2\sum p_i q_i} + 0.01 \mathbf{I}$ or something similar.
Real results (AMASGEN)

• 9 real French bulls among 1827 genotyped, ~50000 SNPs
• Very complex pedigree, simplified graph:
Pedigree-based relationship

Little inbreeding

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,]</td>
<td>1.00</td>
<td>0.51</td>
<td>0.57</td>
<td>0.51</td>
<td>0.26</td>
<td>0.15</td>
<td>0.15</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>[2,]</td>
<td>0.51</td>
<td>1.01</td>
<td>0.30</td>
<td>0.33</td>
<td>0.17</td>
<td>0.17</td>
<td>0.12</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>[3,]</td>
<td>0.57</td>
<td>0.30</td>
<td>1.07</td>
<td>0.30</td>
<td>0.20</td>
<td>0.12</td>
<td>0.18</td>
<td>0.11</td>
<td>0.12</td>
</tr>
<tr>
<td>[4,]</td>
<td>0.51</td>
<td>0.33</td>
<td>0.30</td>
<td>1.01</td>
<td>0.17</td>
<td>0.18</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>[5,]</td>
<td>0.26</td>
<td>0.17</td>
<td>0.20</td>
<td>0.17</td>
<td>1.00</td>
<td>0.56</td>
<td>0.51</td>
<td>0.52</td>
<td>0.53</td>
</tr>
<tr>
<td>[6,]</td>
<td>0.15</td>
<td>0.17</td>
<td>0.12</td>
<td>0.18</td>
<td>0.56</td>
<td>1.06</td>
<td>0.31</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td>[7,]</td>
<td>0.15</td>
<td>0.12</td>
<td>0.18</td>
<td>0.11</td>
<td>0.51</td>
<td>0.31</td>
<td>1.01</td>
<td>0.30</td>
<td>0.29</td>
</tr>
<tr>
<td>[8,]</td>
<td>0.14</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
<td>0.52</td>
<td>0.32</td>
<td>0.30</td>
<td>1.02</td>
<td>0.30</td>
</tr>
<tr>
<td>[9,]</td>
<td>0.14</td>
<td>0.11</td>
<td>0.12</td>
<td>0.11</td>
<td>0.53</td>
<td>0.32</td>
<td>0.29</td>
<td>0.30</td>
<td>1.03</td>
</tr>
</tbody>
</table>

Cousin relationships ~0.125
"first $G$" genomic relationship

Less than 1 in the diagonal

Negative coefficients

$$G = \frac{ZZ'}{2} \sum_{all\,SNPs} p_i (1 - p_i)$$

Relationships among cousins are ~0
GBLUP == GBLUP

• For all matrices of the kind $G = \frac{(M-2P)(M-2P)'}{2\sum p_iq_i}$
  • We don’t need to put the same p’s in the upper and and in the lower part
• Changing allele frequencies in $P$ shifts EBV’s by a constant
• Changing allele frequencies in $\frac{1}{2\sum p_iq_i}$ "scales"
  • But we can compensate through a change in the “genetic variance”

• E.g. $\begin{pmatrix} 1.1 & 0.55 \\ 0.55 & 1.1 \end{pmatrix} 10 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} 11$

• So, if variances are estimated by REML or « corrected » according to $2\sum p_iq_i$ results should be identical : see exercises
IBS relationships at the markers

• $G_{IBS}$ is a genomic relationship matrix based on Identity By State at the markers

• The terms in $G_{IBS}$ are usually described in terms of identities or countings:

\[ G_{IBS_{ij}} = \frac{1}{n} \sum_{m=1}^{n} 2 \left( \sum_{k=1}^{2} \sum_{l=1}^{2} I_{kl} \right) / 4, \]

• where $I_{kl}$ measures the identity across all 4 combinations of alleles
GBLUP == GBLUP with IBS

• $G_{IBS} = \frac{1}{2} G_{0.5} + 11'$ where $G_{0.5}$ is built pretending that $p = 0.5$

• Again, using the right variance components we get the same EBVs
OK, so what should I use?

- For GBLUP it does not matter much, all models are equivalent
- If using REML or Bayesian methods you get the variance components right
- If you use pre-estimated variance components you want to use comparable variances
  - This is a bit tricky but in most cases “default” $\mathbf{G}$ works just fine
- Comparing variance components and $h^2$ also gets tricky
- See Legarra 2016, TPB
OK, so what should I use?

• For SSGBLUP it is essential to have “compatible” genomic and pedigree relationships
• Populations evolve with time, but genotypes came years after pedigree started
• Genomic Predictions are shifted from Pedigree Predictions
• Compatibility is achieved if both relationships refer to the same genetic base:
  • Same average BV at the base
  • Same genetic variance at the base
• Will be presented at SSGBLUP