Genomics

LRS

CGIL

July-Aug 2012

Intro

- Molecular genetics started in 1970's
- Researchers made big promises for the next 5 years, — then the next 5 years, — then the next 5 years,... etc.
- Meuwissen and Goddard began wondering what if those promises came true, better prepare for it.
- Inspired by Ben Hayes course in Guelph, by another Meuwissen, Hayes, Goddard paper, analyzing dense markers across the genome.
- 2006 paper on expected genetic gains in dairy cattle.
- Use of SNP markers took off, but in an uncontrolled manner. Everyone wanted to be part of genomics, even dairy producers, but how?
- Most methods were developed too hastily without proper thought (my opinion).

Study

A slow, thought-out study, to find the best way of utilizing SNP genotypes and gene sequences to improve the accuracy of EBVs of livestock.

Genomics may not be the answer for all species.

Misconception: We do not need data any more. WRONG!!!!
We need more data, and always will.

DNA and Genes

- 3.5 billion base pairs, depending on species
- Thousands of genes (25,000)
  - Where are they?
  - What do they do?
  - How many alleles? frequencies?
- Millions of SNPs, do we need that many?
- Epigenetic effects (PE effects, except they can be transmitted to offspring) - methylization.

Single Locus, Two Alleles

Genotype	Frequency	Value
AA	p^2	10
Aa	$2pq$	9
aa	q^2	-10

Mean

$$\mu_G = p^2(10) \ + \ 2pq(9) \ + \ q^2(-10)$$

Variance

$$\sigma_G^2 = p^2(10)^2 \ + \ 2pq(9)^2 \ + \ q^2(-10)^2 \ - \ \mu_G^2$$

Heritability

$$h^2 \ = \ \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2}$$

Modelling

$$y_{ijk} \; = \; \alpha + g_{ij} + e_{ijk}$$

Least squares estimation, testing

$$\begin{pmatrix} N & n_{11} & n_{12} & n_{22} \\ n_{11} & n_{11} & 0 & 0 \\ n_{12} & 0 & n_{12} & 0 \\ n_{22} & 0 & 0 & n_{22} \end{pmatrix} \begin{pmatrix} \alpha \\ g_{11} \\ g_{12} \\ g_{22} \end{pmatrix} = \begin{pmatrix} y_{...} \\ y_{11.} \\ y_{12.} \\ y_{22.} \end{pmatrix}$$

N should be greater than 3

Simulation Study

$$y_{ijk} = \mu + X_i + H_j + g_k + e_{ijk},$$

where

- μ was 0,
- X_i was a fixed effect of sex (females were +4 and males were -4),
- H_j was a random effect of herd-year (100 herd-years in each generation) with mean 0 and variance 16,
- g_k were the genotype effects as above
- e_{ijk} was a random residual effect with mean 0 and variance 100.

Study

- Progeny were assigned to herd-years randomly
- Females occurred at a frequency of 0.51
- The total phenotypic variance was $(70.25 + 16 + 100) = 186.25$
- Heritability in the broad sense was 0.377
- A base population of 10,000 individuals such that allele frequencies were 0.5
- Six discrete generations, of 10,000 progeny each, were produced through random matings where Mendelian sampling of alleles was followed.
- Phenotypes were generated for animals in generations 5 and 6.

Analysis I

- The same model as used to simulate the data.
- Generation 5 was analyzed, Generation 6 predicted.

Error SS	946,874
R^2	0.50
$\hat{g}_{11}$	5.46
$\hat{g}_{12}$	4.44
$\hat{g}_{22}$	-13.83
$\hat{g}_{11} - \hat{g}_{22}$	19.29
$\hat{g}_{11} - \hat{g}_{12}$	1.02
Females - Males	7.63
Gen 6 SSE	1,078,046

Analysis II

- An animal model was used.

$$y_{ijk} = \mu + X_i + H_j + a_k + e_{ijk}$$

- All relationships among animals were used, genotypes assumed to be unknown.

	Analysis I	II
Error SS	946,874	959,804
R^2	0.50	0.50
$\hat{g}_{11}$	5.46	2.67
$\hat{g}_{12}$	4.44	1.62
$\hat{g}_{22}$	-13.83	-5.91
$\hat{g}_{11} - \hat{g}_{22}$	19.29	8.58
$\hat{g}_{11} - \hat{g}_{12}$	1.02	1.05
Females - Males	7.63	7.64
Gen 6 SSE	1,078,046	1,667415

Analysis III

- Combined model.

$$y_{ijkl} = X_i + H_j + g_k + a_{kl} + e_{ijkl}$$

	Analysis I	II	III
Error SS	946,874	959,804	584,705
R^2	0.50	0.50	.69
$\hat{g}_{11}$	5.46	2.67	5.45
$\hat{g}_{12}$	4.44	1.62	4.41
$\hat{g}_{22}$	-13.83	-5.91	-13.92
$\hat{g}_{11} - \hat{g}_{22}$	19.29	8.58	19.37
$\hat{g}_{11} - \hat{g}_{12}$	1.02	1.05	1.04
Females - Males	7.63	7.64	7.63
Gen 6 SSE	1,078,046	1,667415	

Single Locus, Two alleles

- Combined analysis gave better fit to data, but there should not be any polygenic effects remaining after accounting for genotypes.
- Animal model gave biased estimates of genotype effects, particularly when dominance is present.

Single Locus, More Alleles

- If there are k alleles, then there are $\frac{k(k+1)}{2}$ genotypes.

Alleles	Genotypes
2	3
3	6
4	10
5	15

Two Loci, Unlinked

$$y_{ijklm} = \alpha + g_{1ij} + g_{2kl} + (g_1g_2)ijkl + e_{ijklm}$$

Rank of **X** equals number of interaction terms.
Suppose Locus 1 has 2 alleles and Locus 2 has 3 alleles, then the number of interaction terms is 3 × 6 = 18.

	g_{111}	g_{112}	g_{122}	Row Effect
g_{211}	1	0	-1	-3
g_{212}	3	2	-2	-5
g_{213}	1	1	0	-1
g_{222}	2	-3	1	0
g_{223}	1	5	-7	2
g_{233}	2	3	-1	7
Col Effect	20	14	-15	

Two Loci, Unlinked

$$\begin{aligned}
\sigma^2_{10} &= \text{Additive} \\
\sigma^2_{01} &= \text{Dominance} \\
\sigma^2_{11} &= \text{Add x Dom} \\
\sigma^2_{20} &= \text{Add x Add} \\
\sigma^2_{02} &= \text{Dom x Dom}
\end{aligned}$$

Three Loci

$$\begin{aligned} y &= \alpha + \sum_{i=1}^{3} g_i \\ &\quad +(g_1 g_2) + (g_1 g_3) + (g_2 g_3) \\ &\quad +(g_1 g_2 g_3) + e \end{aligned}$$

$$a = \sum_{i=1}^{3} g_i$$

No dominance effects.
Need to genotype more individuals.

25,000 Loci

- 312,487,500 possible two-way interactions
- 2.6 trillion three-way interactions
- 4, 5, 6, ... 25000 - way interactions too.
- Impossible to estimate all of these, unless we make many assumptions.

SNPs

- SNPs, biallelic, millions of them
- Close to or inside genes
- Used for
  - Locating genes of importance
  - Estimating effects of small segments of DNA, to give EBV
  - Studying pathways of biology in different organs through genes
- Genomic selection of breeding individuals at an earlier age, more accurately than PA

Association Tests

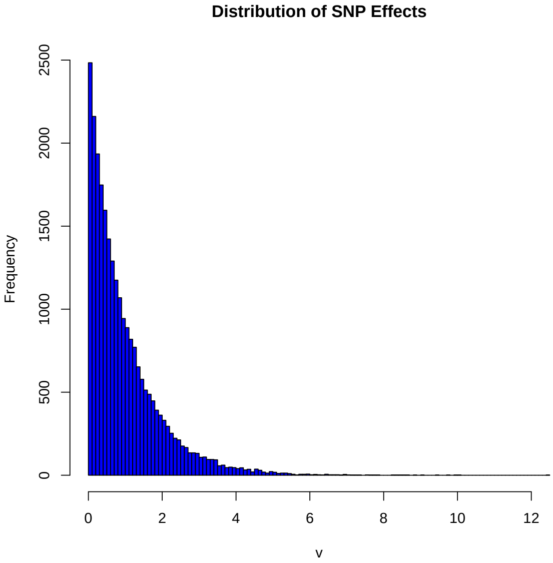
$$\mathbf{y} = Fixed + Random + \mathbf{a} + \mathbf{W}_1(SNP_1) + \mathbf{e}$$

$$(SNP_1) \ = \ \begin{pmatrix} s_{11} \\ s_{12} \\ s_{22} \end{pmatrix} \ = \ b_1 p$$

where $p == 1$ if SNP genotype s_{11}, $p = 0$ if SNP genotype s_{12}, or $p = 1$ if SNP genotype s_{22}. $\mathbf{W}$ has order $N \times 3$

- **a** is the additive polygenic effect for everything not accounted for by the SNP genotypes, **A** used.
- One SNP is analyzed at a time, to find the most significant.
- One has to adjust criteria for significance levels - because there are so many SNPs.

Distribution of SNP effects



Generating SNP Effects

```
nsnp = 50
v = round(rgamma(nsnp,1,scale=1)*100)/100
snpfx[ ,1] = v   # genotype 11
snpfx[ ,2] = 0   # genotype 12
snpfx[ ,3] = -v  # genotype 22
```

Frequencies of SNP Alleles

```
  fp = round(runif(nsnp,0.11,0.89)*100)/100
  fq = 1 - fp
#  Frequencies of genotypes (all SNPs)
  fgg = cbind((fp*fp),(2*fp*fq),(fq*fq))
  kgg = c(1,2,3)
```

Mean and Variance

```
for(isnp in 1:nsnp){
 avesnp[isnp] = sum(fgg[isnp, ]*snpfx[isnp, ])
vsnp[isnp] = sum(fgg[isnp, ]*snpfx[isnp, ]*snpfx[isnp, ])
  - avesnp[isnp]*avesnp[isnp]
}
mu = sum(avesnp)
 -3.4
vv = sum(vsnp)
 45.47
```

Base Animals

```
for(iam in 1:nbase){
  for(isnp in 1:nsnp){
  q = fgg[isnp, ]
  jgg = sample(kgg,1,prob=q)
  atbv[iam]=atbv[iam]+snpfx[isnp,jgg]
  asnp[iam,isnp]=jgg
 }
 }
```

Progeny

```
  for(iam in (nbase+1):nam){
# determine parents, sample(aid,2,replace=FALSE)
  for(isnp in 1:nsnp){
    ksgg = asnp[ks,isnp]; kdgg=asnp[kd,isnp]
    als = gtoa[ksgg,1]; if(runif(1)>0.5)als=gtoa[ksgg,2]
    ald = gtoa[kdgg,1]; if(runif(1)>0.5)ald=gtoa[kdgg,2]
    kagg = atog[als,ald]
    atbv[iam]=atbv[iam]+snpfx[isnp,kagg]
    asnp[iam,isnp] = kagg  }
    }
```

Create Phenotypes

```
agegp = c(1:4)
agefx = c(100,120,130,135)
SDe=sqrt(140)
obs=round(agefx(age) + atbv + rnorm(nam,0,SDe))
kattle=data.frame(aid,sid,did,age,obs,atbv)
```

Analysis I

$$y = \text{age} \; + \; \text{animal} \; + \; e$$

- Usual animal model, using $\mathbf{A}^{-1}$
- Ignore genotypes
- All 20 animals used
- $\sigma_e^2/\sigma_a^2 \; = \; 140/45.47$
- Get correlation of EBV and TBV, also SSE

Analysis II

$$y = \text{age} \ + \ a \ + \ b * gg \ + \ e$$

- *gg* is SNP genotype, coded as (-1,0,1), covariate
- One SNP at a time, i.e. 50 analyses
- Only animals with genotypes (8) and full **A**.
- Test *b* for each SNP

Analysis III

$$DYD \; = \; \mu \; + \; \sum (b_i * gg_i) \; + \; e$$

- *DYD* is daughter yield deviation, or EBV with accuracy of 0.90 or better
- All 50 SNPs used at one time
- 8 observations, 50 SNPs, more parameters to estimate than there are observations
- Need a validation data set

Form G Matrix, Analysis IV

$$\mathbf{G} \;=\; \mathbf{TT}'/(\sum 2p_iq_i)$$

where $\mathbf{T}$ is $q \times m$, animals by SNPs,

$$Var(\mathbf{a}) \;=\; \mathbf{G}\sigma_a^2$$

$$DYD \;=\; \mu \;+\; a \;+\; e$$

- Only animals with genotypes
- **G** has to be inverted
- gEBV produced, scaling

Two Steps, Analysis V

- How to use gEBV to change EBV of all animals
- Selection index?
- Weight EBV and gEBV by accuracies
- Should this be done?
- Are current procedures appropriate?

One Step Method, Analysis VI

Misztal, Legarra, Aguilar (2009)

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e}$$

where

$$\begin{aligned} \mathbf{Z} &= \begin{pmatrix} \mathbf{Z}_1 & \mathbf{Z}_2 \end{pmatrix} \\ \mathbf{a} &= \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{pmatrix} \end{aligned}$$

for 1 being animals not genotyped and 2 denoting animals that have been genotyped. Then

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$$

One Step, MME

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{H}^{-1}\alpha \end{pmatrix} \begin{pmatrix} \mathbf{b} \\ \mathbf{a} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{pmatrix}$$

where

$$\mathbf{H}^{-1} = \begin{pmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22}+\mathbf{G}^{-1}-\mathbf{A}_{22}^{-1} \end{pmatrix}$$

Ducrocq and Legarra, 2011

$$\begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1^* + \mathbf{d}_1 \\ \mathbf{a}_2^* + \mathbf{d}_2 \end{pmatrix}$$

because $\mathbf{a}_1$ animals are not genotyped, then

$$\mathbf{d}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{d}_2$$

Ducrocq and Legarra, 2011

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_1 & \mathbf{X}'\mathbf{Z}_2 & \mathbf{0} \\ \mathbf{Z}'{}_1\mathbf{X} & \mathbf{Z}'{}_1\mathbf{Z}_1 + \alpha\mathbf{A}^{11} & \alpha\mathbf{A}^{12} & \mathbf{0} \\ \mathbf{Z}'{}_2\mathbf{X} & \alpha\mathbf{A}^{21} & \mathbf{Z}'{}_2\mathbf{Z}_2 + \alpha\mathbf{A}^{22} & -\alpha\mathbf{A}_{22}^{-1} \\ \mathbf{0} & \mathbf{0} & -\alpha\mathbf{A}_{22}^{-1} & \alpha(\mathbf{A}_{22}^{-1} + [\mathbf{G} - \mathbf{A}_{22}]^{-1}) \end{pmatrix}$$

$$\begin{pmatrix} \mathbf{b} \\ \mathbf{a}_1 \\ \mathbf{a}_2 \\ \mathbf{d}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'{}_1\mathbf{y} \\ \mathbf{Z}'{}_2\mathbf{y} \\ \mathbf{0} \end{pmatrix}$$

Iteration Strategy

1. Start with $\mathbf{d}_2 = \mathbf{0}$
2. Solve

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_1 & \mathbf{X}'\mathbf{Z}_2 \\ \mathbf{Z'}_1\mathbf{X} & \mathbf{Z'}_1\mathbf{Z}_1 + \alpha\mathbf{A}^{11} & \alpha\mathbf{A}^{12} \\ \mathbf{Z'}_2\mathbf{X} & \alpha\mathbf{A}^{21} & \mathbf{Z'}_2\mathbf{Z}_2 + \alpha\mathbf{A}^{22} \end{pmatrix}$$

$$\begin{pmatrix} \mathbf{b} \\ \mathbf{a}_1 \\ \mathbf{a}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z'}_1\mathbf{y} \\ \mathbf{Z'}_2\mathbf{y} + \alpha\mathbf{A}_{22}^{-1}\mathbf{d}_2 \end{pmatrix}$$

3. Solve

$$(\mathbf{A}_{22}^{-1} + [\mathbf{G} - \mathbf{A}_{22}]^{-1})\mathbf{d}_2 \;=\; \mathbf{A}_{22}^{-1}\mathbf{a}_2$$

4. Iterate steps 2 and 3 to convergence.

Analysis VII

There are many genes(SNPs) with small effects (millions)

- 10,000 individuals generated (five generations), 5000 SNPs each
- Heritability = 0.25
- Both sexes have a phenotype
- Genetic breeding value = sum of all SNP effects

$$\mathbf{y} = Fixed + Random + \mathbf{a} + \sum_{i=1}^{m} \mathbf{W}_i(SNP_i) + \mathbf{e}$$

for m going from 1 to 70.

$$EBV = \hat{\mathbf{a}} + \sum_{i=1}^{m} \mathbf{W}_i(SNP_i)$$

Results

SNPs	Residual Var.	diff.
0	763.22	-
10	736.87	26.35
20	716.25	20.62
30	698.71	17.54
40	683.36	15.35
50	668.72	14.64
60	655.45	13.27
70	643.33	12.12

15.7% Reduction.

Considerations

- All animals need genotypes
- Instead of imputation we need better segregation analysis (Kerr and Kinghorn) combines data and genotypes in Bayesian method.

Ramblings

- SNPs have 2 alleles, genes have more than 2, most likely, thus SNP marker effects can be unstable depending on allele frequencies.
- Crossbreds, not sure if SNPs will work well or not.
- More simulation studies needed to compare methodologies, not to come up with more new method.