Bases for Genomic Prediction

Andres Legarra
Council on Dairy Cattle Breeding, Bowie, MD
"Ad honorem" Adjunct Assistant Professor in Animal Genetics and Genomics, University of Georgia

1. History
2. Markers
3. SNP-BLUP
4. G-BLUP
5. SS-GBLUP
6. Details & Horror Stories
7. Method LR

- Pre-2005: much is said about markers and QTL but little is done that has practical results.
  - genotyping markers (microsatellites) is time-consuming and expensive. Technologies are refined thanks, in part, to the Human Genome Project and the like.
  - Around this time: a "cattle" consortium is created to join forces and create a common SNP chip.
- 2007:
  - VanRaden presents at Interbull the concepts of genomic relationship (intuited by many people but never well formalized until his presentation)
  - at the QTLMAS meeting in Toulouse, EAAP and other sites, first genomic evaluation results are presented, still very experimental and with much reduced datasets
- 2008:
  - in April, USDA launches the first internal genomic evaluation and at the end of the year it is official.
  - VanRaden publishes his paper, full of ideas, highly cited but little read. The same year, the official methodology is presented in detail (VanRaden et al 2009).
  - It is quickly understood that the proposed methods cannot be applied in the case "some animals are not genotyped" -> need for SSGBLUP.
- 2009:
  - in January we (Legarra-Aguilar-Misztal) sent the SSGBLUP paper to the Journal of Dairy Science. The idea is well received.
  - In August it is presented at Interbull. Ole Christensen (U of Aarhus) presents the same developments done in parallel (and in a more elegant way).
- 2010 - 2014
  - Many skeptics but nobody finds something better
- 2014 -
  - Generally accepted. Refinements and computational strategies, but the basic concept remains the same.

MARKERS

Data files

```
64546020 1980 6 17 15 13 4 1 2 0 0 0  1676.8800
64546020 1981 7 18 16 12 5 1 2 0 0 0  1433.6090
64546020 1982 8 17 15 14 5 1 2 0 0 0   769.2500
64546020 1980 6 17 15 12 4 1 2 0 0 0  1466.4200
64546020 1981 7 18 16 13 5 1 2 0 0 0  1474.4940
64546020 1982 8 17 16 14 5 1 2 0 0 0  1523.0290
64546020 1980 6 17 15 12 4 1 2 0 0 0  1601.2290
64546020 1981 7 17 16 12 5 1 2 0 0 0  1205.8100
64546020 1982 8 17 16 14 5 1 2 0 0 0   676.5150
64546020 1980 5 17 15 13 3 1 2 0 0 0  2122.1930
64546020 1981 6 17 15 12 4 1 2 0 0 0  2227.5940
64546020 1982 7 17 16 14 5 1 2 0 0 0  1593.4090
64546020 1980 5 17 16 11 3 1 2 0 0 0  2132.2250
64546020 1981 6 17 16 13 4 1 2 0 0 0  2100.5200
64546020 1982 7 17 16 12 5 1 2 0 0 0  1792.2250
64546020 1983 8 17 15 12 5 1 2 0 0 0  1492.0900
64546020 1984 9 17 15 11 5 1 2 0 0 0  1607.3500
64546020 1985 10 17 15 12 5 1 2 0 0 0  1534.3350
64546020 1986 11 18 17 15 6 1 2 0 0 0   958.1200
```


Pedigree files

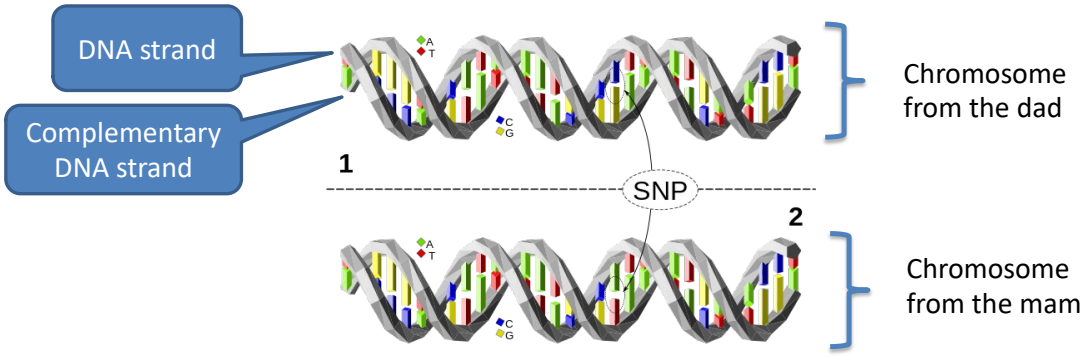
```
00000700640031;00000000000000;00000700620012;1964;2
00000700640032;00000700620045;00000700600138;1964;2
00000700640033;00000700630065;00000700540069;1964;2
00000700640034;00000000000000;00000700580089;1964;2
00000700640035;00000000000000;00000700590106;1964;2
00000700640036;00000700630065;00000700550017;1964;2
00000700650001;00000700620047;00000700610007;1965;2
00000700650002;00000702630050;00000700560023;1965;2
00000700650003;00000700620047;00000700600125;1965;2
00000700650004;00000700620047;00000700620027;1965;2
```

Marker files?

```
64000670990546 120120202102111210122210200 0
   45214790003 121111202111010210222220200 1
   45214790004 221110201101021200222211211 0
   45199680012 210111102011011210122201211 0
   45307160107 121210202002022200222211211 0
   45199690008 220210202001022211222210211 1
64000249040705 121200202001022210122201211 0
   45189980105 120110202101111220022200200 0
64000249030710 221110201101112211222201222 1
   45214790006 221200202000021211122210101 1
   45199680014 122201101101012222222111111 0
   45214780461 221210201102022201122221111 1
   45253180017 110111102000202221222210222 2
64000311010387 122200202001021210122201211 0
```


What are SNPs

- SNPs: https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism



The upper DNA molecule differs from the lower DNA molecule at a single base-pair location (a C/A polymorphism)

Genotype files

- SNP files come from some machines
- In some obscure format
- We need to understand the format to understand what we do later
- Some people deal with raw files, some people do not


```
[Header]
GSGT Version    1.9.4
Processing Date 3/16/2012 9:11 AM
C                    50_B
N
T
N
T
[Data]
Sample ID       Sample Name     SNP Name        Allele1 - Top   Allele2 - Top   GC Score
ES140000270478  PLACA_CIC_12_96 250506CS3900065000002_1238.1    G       G       0.8932
ES140000270478  PLACA_CIC_12_96 250506CS3900140500001_312.1     A       G       0.7341
ES140000270478  PLACA_CIC_12_96 250506CS3900176800001_906.1     A       G       0.7532
ES140000270478  PLACA_CIC_12_96 250506CS3900211600001_1041.1    A       A       0.9674
ES140000270478  PLACA_CIC_12_96 250506CS3900218700001_1294.1    G       G       0.8178
ES140000270478  PLACA_CIC_12_96 250506CS3900283200001_442.1     C       C       0.6684
ES140000270478  PLACA_CIC_12_96 250506CS3900371000001_1255.1    G       G       0.4565
ES140000270478  PLACA_CIC_12_96 250506CS3900386000001_696.1     A       A       0.4258
ES140000270478  PLACA_CIC_12_96 250506CS3900414400001_1178.1    G       G       0.8690
ES140000270478  PLACA_CIC_12_96 250506CS3900435700001_1658.1    A       A       0.5153
ES140000270478  PLACA_CIC_12_96 250506CS3900464100001_519.1     A       G       0.8116
ES140000270478  PLACA_CIC_12_96 250506CS3900487100001_1521.1    A       G       0.7448
ES140000270478  PLACA_CIC_12_96 250506CS3900539000001_471.1     G       G       0.5248
ES140000270478  PLACA_CIC_12_96 250506CS3901012300001_913.1     A       A       0.7413
ES140000270478  PLACA_CIC_12_96 250506CS3901300500001_1084.1    G       G       0.7990
ES140000270478  PLACA_CIC_12_96 CL635241_413.1  A       A       0.8176
ES140000270478  PLACA_CIC_12_96 CL635750_128.1  A       G       0.7978
ES140000270478  PLACA_CIC_12_96 CL635944_160.1  A       G       0.7283
```

Animal

SNP name

Genotype in nucleotides

SNP name

Animal

Genotype in A/B format

10:11 AM
0_v2_C.bpm

[Data]

SNP Name	Sample ID	Allele1 - Forward	Allele2 - Forward	Allele1 - Top	Allele2 - Top	Allele1 - AB	Allele2 - AB	GC Score	X	Y
ARS-BFGL-BAC-10172	USA201811	G	G	G	B	B	0.9506	0.012	1.036	
ARS-BFGL-BAC-1020	USA201811	G	G	G	B	B	0.9673	0.005	0.652	
ARS-BFGL-BAC-10245	USA201811	C	C	G	B	B	0.7579	0.092	1.417	
ARS-BFGL-BAC-10345	USA201811	A	A	A	A	A	0.9276	1.143	0.008	
ARS-BFGL-BAC-10365	USA201811	G	G	C	B	B	0.5335	0.004	0.862	
ARS-BFGL-BAC-10375	USA201811	A	G	A	A	B	0.9567	0.478	0.581	
ARS-BFGL-BAC-10591	USA201811	A	G	A	A	B	0.9003	0.386	0.473	
ARS-BFGL-BAC-10867	USA201811	G	G	C	A	A	0.9434	0.776	0.004	
ARS-BFGL-BAC-10919	USA201811	A	A	A	A	A	0.8526	1.232	0.036	
ARS-BFGL-BAC-10951	USA201811	T	T	A	A	A	0.5140	0.539	0.017	
ARS-BFGL-BAC-10952	USA201811	A	A	A	A	A	0.9512	0.987	0.030	
ARS-BFGL-BAC-10960	USA201811	G	G	G	B	B	0.9528	0.018	0.826	
ARS-BFGL-BAC-10972	USA201811	G	C	C	G	A	B	0.8759	0.917	0.743
ARS-BFGL-BAC-10975	USA201811	A	G	A	G	A	B	0.8142	0.979	0.739
ARS-BFGL-BAC-10986	USA201811	G	G	C	C	B	B	0.9309	0.055	0.731
ARS-BFGL-BAC-10993	USA201811	C	C	G	G	B	B	0.9014	0.023	1.094
ARS-BFGL-BAC-11000	USA201811	T	T	A	A	A	A	0.9686	0.561	0.013
ARS-BFGL-BAC-11003	USA201811	T	T	A	A	A	A	0.9215	1.171	0.040
ARS-BFGL-BAC-11007	USA201811	T	C	A	G	A	B	0.9454	0.884	0.675
ARS-BFGL-BAC-11025	USA201811	G	G	C	C	B	B	0.9082	0.015	0.740
ARS-BFGL-BAC-11028	USA201811	A	G	A	G	A	B	0.9678	0.182	0.288
ARS-BFGL-BAC-11034	USA201811	T	C	A	G	A	B	0.9509	0.566	0.592
ARS-BFGL-BAC-11039	USA201811	C	C	G	G	B	B	0.9658	0.000	0.889
ARS-BFGL-BAC-11042	USA201811	A	G	A	G	A	B	0.8506	0.947	0.786
ARS-BFGL-BAC-11044	USA201811	T	C	A	G	A	B	0.9654	0.726	0.689
ARS-BFGL-BAC-11047	USA201811	T	T	A	A	A	A	0.9465	0.973	0.015

SNP name

Chromosome number

Position in base pairs

```
index,Nom,OAR,Num,Pos
1,250506CS3900065000002_1238.1,15,95,5825554
2,250506CS3900140500001_312.1,23,471,26446680
3,250506CS3900176800001_906.1,7,1828,81627347
4,250506CS3900211600001_1041.1,16,919,41632053
5,250506CS3900218700001_1294.1,2,3311,149375044
6,250506CS3900283200001_442.1,1,4056,188745186
7,250506CS3900371000001_1255.1,11,657,35486157
8,250506CS3900386000001_696.1,16,1391,62983985
9,250506CS3900414400001_1178.1,1,2238,103373031
10,250506CS3900435700001_1658.1,12,976,44985453
11,250506CS3900464100001_519.1,1,1859,85681719
12,250506CS3900487100001_1521.1,14,21,1046097
13,250506CS3900539000001_471.1,27,1189,101575221
14,250506CS3901012300001_913.1,2,2240,100935467
15,250506CS3901300500001_1084.1,7,2015,89446225
16,CL635241_413.1,3,4089,181937734
17,CL635750_128.1,3,5009,223456572
18,CL635944_160.1,6,2374,107677235
19,Contig35697_5761.1,6,397,18930545
20,CR_594.1,27,597,51062613
21,CR_816.1,27,595,51062391
22,CytB_1406.1,3,4592,204780199
23,CytB_1505.1,3,4593,204780298
```

Map file

This format is very uncomfortable
It is easier to have 1 line/animal

```
[Header]
GSGT V
Proces
Conten
Num SN
Total SNPs       54241
Num Samples      36
Total Samples    36
[Data]
Sample ID        Sample Name      SNP Name         Allele1 - Top   Allele2 - Top   GC Score
ES140000270478   PLACA_CIC_12_96  250506CS3900065000002_1238.1     G       G       0.8932
ES140000270478   PLACA_CIC_12_96  250506CS3900140500001_312.1      A       G       0.7341
ES140000270478   PLACA_CIC_12_96  250506CS3900176800001_906.1      A       G       0.7532
ES140000270478   PLACA_CIC_12_96  250506CS3900211600001_1041.1     A       A       0.9674
ES140000270478   PLACA_CIC_12_96  250506CS3900218700001_1294.1     G       G       0.8178
ES140000270478   PLACA_CIC_12_96  250506CS3900283200001_442.1      C       C       0.6684
ES140000270478   PLACA_CIC_12_96  250506CS3900371000001_1255.1     G       G       0.4565
ES140000270478   PLACA_CIC_12_96  250506CS3900386000001_696.1      A       A       0.4258
ES140000270478   PLACA_CIC_12_96  250506CS3900414400001_1178.1     G       G       0.8690
ES140000270478   PLACA_CIC_12_96  250506CS3900435700001_1658.1     A       A       0.5153
ES140000270478   PLACA_CIC_12_96  250506CS3900464100001_519.1      A       G       0.8116
ES140000270478   PLACA_CIC_12_96  250506CS3900487100001_1521.1     A       G       0.7448
ES140000270478   PLACA_CIC_12_96  250506CS3900539000001_471.1      G       G       0.5248
ES140000270478   PLACA_CIC_12_96  250506CS3901012300001_913.1      A       A       0.7413
ES140000270478   PLACA_CIC_12_96  250506CS3901300500001_1084.1     G       G       0.7990
ES140000270478   PLACA_CIC_12_96  CL635241_413.1   A       A       0.8176
ES140000270478   PLACA_CIC_12_96  CL635750_128.1   A       G       0.7978
ES140000270478   PLACA_CIC_12_96  CL635944_160.1   A       G       0.7283
```


- 1 line/animal

```
ES1400NAB40571 G G G G A A A C . . A G
ES1400NAB40573 G G G G G G A C G G A G
ES1400NAB40574 A G G G A G A C G G A A
ES1400NAB40159 G G G G A G A C G G A A
ES1400NAB40528 A G A G A G C C A G A A
ES1500VI492705 G G A G G G A C G G A G
ES1500SSA40533 A G G G A G C C G G A A
```

Marker 1 Marker 2

- Animal breeders and computers don't like text, prefer numbers
- At each marker locus, there are only two possible alleles, for instance:
  - For marker 1 this could be A / C
  - For marker 2 this could be A / G
- Then we choose one of those markers as the reference one

Gene content

- For instance if there are two nucleotides (A/C) and C is the reference:
- 0 means AA
- 1 means AC or CA
- 2 CC
- 5 missing

- For another loci, the reference allele might be something else
- This way of coding is known as "gene content"
- One column (and not 2) per marker

Gene content

- For instance if there are two nucleotides (A/C) and A is the reference:
- 2 means AA
- 1 means AC or CA
- 0 CC
- 5 missing

- For another loci, the reference allele might be something else
- This way of coding is known as “gene content”
- One column (and not 2) per marker

The reference allele can vary across loci. For instance, consider the same animal

Missing !!

ES1400NAB40571 G G G G A A A C . . A G

And consider that the reference alleles for each of the 6 markers are (G,G,A,C,G,A). Using these reference alleles would give

ES1400NAB40571 222151

Missing !!

Final genotype file

```
64000670990546 120120202102111210122210200 0
   45214790003 121111202111010210222220200 1
   45214790004 221110201101021200222211211 0
   45199680012 210111102011011210122201211 0
   45307160107 121210202002022200222211211 0
   45199690008 220210202001022211222210211 1
64000249040705 121200202001022210122201211 0
   45189980105 120110202101111220022200200 0
64000249030710 221110201101112211222201222 1
   45214790006 221200202000021211122210101 1
   45199680014 122201101101012222222111111 0
   45214780461 221210201102022201122221111
   45253180017 110111102000202221222210222
64000311010387 122200202001021210122201211 0
```

Animal

Long “row” with thousands of markers

Final genotype file

How can I read this?

```
640006
   452
   45214790004 221110201101
   45199680012 210111102011 2101222012110
   45307160107 121210202002022200222211211 0
   45199690008 220210202001022211222210211 1
64000249040705 121200202001022210122201211 0
   45189980105 120110202101111220022200200 0
64000249030710 221110201101112211222201222 1
   45214790006 221200202000021211122210101 1
   45199680014 122201101101012222222111111 0
   45214780461 221210201102022201122221111 1
   45253180017 110111102000202221222210222 2
64000311010387 122200202001021210122201211 0
```

How do we edit these files?

SNP chips:

- PLINK ! but you are limited by what plink can do
- Often you need extra editing
- Efficient: Fortran, awk, bash scripts
- Less efficient (usually usable): Python, R
- Learn some programming

Fortran:

```
read(1,'(a14,1x,60000i1)') id,genotype(:)
```

Awk:

```
split($2,genotype,"")
```

Python:

```
for line in fhand:
      idd , genotype = line.split()
            for j,m in enumerate(genotype):
```

Keep track

- If you do these manipulations, you need to keep track of:
- SNP names
- Reference alleles at each locus

- If you mix files make sure that you're working with the same markers and reference alleles!!

- What you see in genotype files
- Minimum quality control

What you find in genotype files

- « call rate » is the percentage of observed genotypes:
  - per animal (per row)
  - per marker (per column)
- In other words, the number of “5”s
- If call rate animal <95% the genotype of the animal is rejected (delete line)
- If call rate marker <95% the column of the marker is deleted

Allele frequency

- The allele frequency p is simply the frequency of the reference allele. For instance consider

```
ES1400NAB40571 G G
ES1400NAB40573 G G
ES1400NAB40574 A G
ES1400NAB40159 G G
ES1400NAB40528 A G
ES1500VI492705 G G
ES1500SSA40533 A A
```

- If the reference allele is G, we have 10G against 4A: $p = \frac{10}{14} \approx 0.71$, and the frequence of allele A is $q = 1 - p \approx 0.29$.

Allele frequency

- When we use integer codes, it is very easy

```
ES1400NAB40571 2
ES1400NAB40573 2
ES1400NAB40574 1
ES1400NAB40159 2
ES1400NAB40528 1
ES1500VI492705 2
ES1500SSA40533 0
```

- p is obtained summing the : $p = \frac{10}{2\times7} \approx 0.71$, and $q = 1 - p \approx 0.29$.

Minor allele frequency

- MAF is the lowest of the two allele frequencies. For instance if the two alleles are A/G
- $p = freq(A); q = 1 - p = freq(G)$
- $MAF = min(p, q)$
- Why is MAF important?

Minor allele frequency

- $MAF = min(p, q)$
- Why is this important?
- A fixed marker ($p = 0$ or $p = 1$) gives no information
- An almost-fixed marker ($p = 0.0001$ or $p = 0.9999$) gives almost no information
- Some applications use $1/p$
- But $\frac{1}{0.000001} = 10^6$, may lead to overflow !!
- So, people delete markers with MAF<0.01 or <0.05
- For prediction and GWAS it does not make much difference
- For sequence analysis with *de novo* variants it makes a difference

How do we compute these things?

Assume that genotypes are stored as 0/1/2 in matrix Z

- `cr_animal(i)=sum(Z(i,:)/=5)/nsnp`
- `cr_marker(i)=sum(Z(:,i)/=5)/nsnp`

Assume no missing values

- `p(i)=sum(Z(:,i))/(2*nanim)`
- `maf(i)=minval((/p(i),1-p(i)/))`

Hardy-Weinberg Equilibrium

- If animals reproduce at random we expect to find HW proportions of genotypes:

$$p^2, 2pq, q^2$$

- We can use a Chi-2 test to test this, but
  - Does HWE equilibrium this hold?
  - Only approximately
  - At each generation p changes a little bit, so it does not hold across all generations
  - Also, animals do not mate at random

Hardy-Weinberg Equilibrium

Rule of thumb used by AIPL (Wiggans 2011):

- Number of heterozygotes should not deviate too much
- Delete marker if $\left|\frac{n\ of\ heterozygotes}{n} - 2pq\right| > 0.15$

Crosses

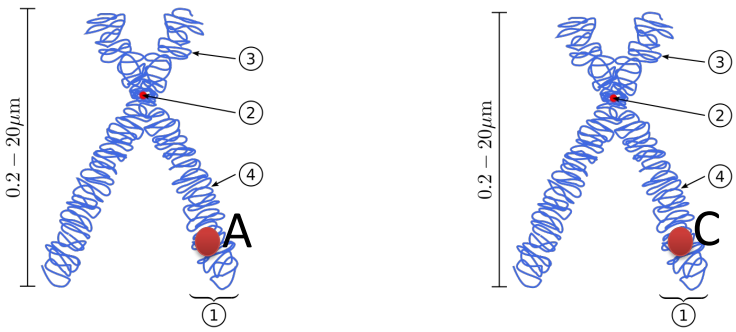
- In crosses you don't expect to have HWE
- Imagine F1 sows from Landrace boars (with allele frequency p^L) and Yorkshire sows (with allele frequency p^Y)
- Then the genotype proportions are

$$(p^Y p^L, p^Y q^L + q^Y p^L, q^Y q^L)$$

- (Why ?)

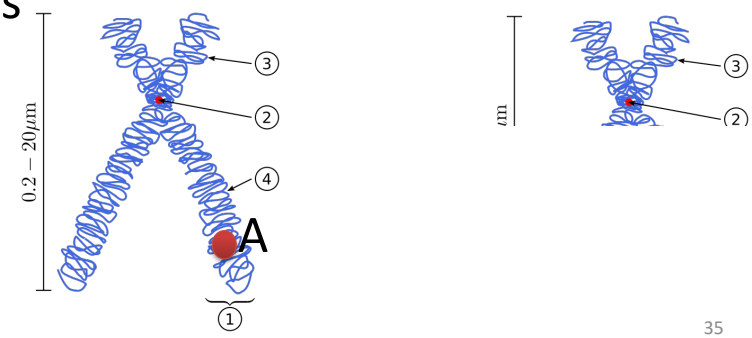
Sex chromosomes 1

- XX (females, mammals) or ZZ (males, birds)
- Two alleles (one from the mam, one from the dad)



Sex chromosomes 2

- XY (males, mammals) or ZW (females, birds)
- <u>One</u> allele (from the mam in mammals, from the dad in birds)
- This is weird and often we don't use these chromosomes



Sex chromosomes 3

- Use of sex chromosomes in prediction is complicated (US dairy does, though)
  - VanRaden et al.. J Dairy Sci. 2009;92:16–24.
  - Druet & Legarra. (2020) Gen Sel Evol , 52(1), 1-17.
- in the course we assume all are autosomes

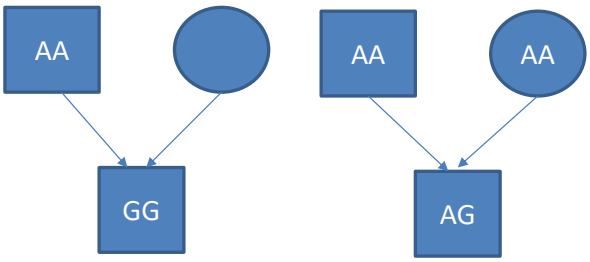
Un mapped markers

- Markers reside in chromosomes
- The position of some markers is still unknown !
- This is reported as “chromosome 0”
- It is better to abandon these markers
- For instance http://www.livestockgenomics.csiro.au/sheep/oar3.1.php :


```
##gff-version3
nohit	SNP50	SNP	.	.	.	.	.	ID=CytB_131;Note=OARv3.1::::pastOARv1.0position:Chr0:0;Alias=CytB_131.1
nohit	SNP50	SNP	.	.	.	.	.	ID=CytB_1406;Note=OARv3.1::::pastOARv1.0position:Chr0:0;Alias=CytB_1406.1
nohit	SNP50	SNP	.	.	.	.	.	ID=CytB_1505;Note=OARv3.1::::pastOARv1.0position:Chr0:0;Alias=CytB_1505.1
nohit	SNP50	SNP	.	.	.	.	.	ID=CytB_1745;Note=OARv3.1::::pastOARv1.0position:Chr0:0;Alias=CytB_1745.1
nohit	SNP50	SNP	.	.	.	.	.	ID=DU287575_503;Note=OARv3.1::::pastOARv1.0position:Chr2:31209299;Alias=DU287575_503.1
nohit	SNP50	SNP	.	.	.	.	.	ID=DU369175_467;Note=OARv3.1::::pastOARv1.0position:Chr4:78053478;Alias=DU369175_467.1
nohit	SNP50	SNP	.	.	.	.	.	ID=DU407749_370;Note=OARv3.1::::pastOARv1.0position:Chr5:80350180;Alias=DU407749_370.1
nohit	SNP50	SNP	.	.	.	.	.	ID=DU415336_399;Note=OARv3.1::::pastOARv1.0position:Chr8:96150336;Alias=DU415336_399.1
nohit	SNP50	SNP	.	.	.	.	.	ID=DU420655_308;Note=OARv3.1::::pastOARv1.0position:Chr12:57781103;Alias=DU420655_308.1
nohit	SNP50	SNP	.	.	.	.	.	ID=DU428219_359;Note=OARv3.1::::pastOARv1.0position:Chr6:113162488;Alias=DU428219_359.1
nohit	SNP50	SNP	.	.	.	.	.	ID=DU439696_403;Note=OARv3.1::::pastOARv1.0position:ChrX:37790463;Alias=DU439696_403.1
```

Mendelian conflicts

- In absence of mutation (which is rare) this kind of things cannot happen:



Mendelian conflicts

- If a marker is seen in many Mendelian conflicts,
  - possibly the genotyping of the marker is wrong and the marker is deleted
- If an animal is seen in many Mendelian conflicts,
  - Possibly there is a misidentification in animal or in pedigree
- You may try to find this animals' parent:
  - Seekparent.f90

Duplicate genotypes

- Two animals should not have identical SNPs unless they are clones or monozygotic twins
- This is unusual…
- If not clones, duplicated genotypes come from mislabeling: the DNA sample of the same animal has been given two different names

Two markers !!

- (or one marker and one QTL)

Linkage disequilibrium

- « Gametic phase disequilibrium »

Statistical association between alleles at two loci in the same chromosome

  - Loci : places
  - Alleles: alternative forms of a gene (A,B,0)
  - Phase: notion of being in the same chromosome (of a pair) or coming from same origin (sire or dam)

Biallelic case

- Assume we genotype 5 individuals, thus 10 chromosomes (and that we know the phase)
- Now we compute allelic frequencies

AB
AB
ab
aB
ab
ab
Ab
AB
Ab
AB

Biallelic case

p(A)=0.6

p(B)=0.5

if independent, p(AB)=0.3,p(ab)=0.2

The expected proportions are:

	A	a
B	0.3	0.2
b	0.3	0.2

Biallelic case

p(A)=0.6
p(B)=0.5
in reality:

	A	a
B	0.4	0.2
b	0.1	0.3

vs. **expected**

	A	a
B	0.3	0.2
b	0.3	0.2

More AB & ab than expected !!
This is **linkage disequilibrium**

Linkage disequilibrium

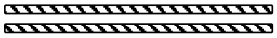
- Is a *statistical* concept
- Describes not-random association of two loci
  - Nothing more, so, why is it useful?
- Two loci in LD *most often* are (very) close
  - This is because LD breaks down with recombination
- Linkage disequilibrium of two loci decays *on average* with the distance
- Hence it serves to map genes

Where does it come from?

- Because chromosomes are transmitted together
  - Within known families (« linkage analysis »)
  - Within the history of a population (« populational linkage disequilibrium » or « linkage disequilibrium » in short)

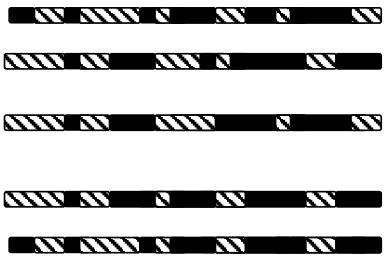
- <u>This distinction is rather artificial</u>
  - Remember: a population *is* a very old, large family

Populational linkage disequilibrium

- Assume we mix two populations (say Churra and Merino)
- Or, that Adam was 
  - and Eve 
  - The first generation is an F1
  - Then animals are mixed at random
- What do we get after many generations?

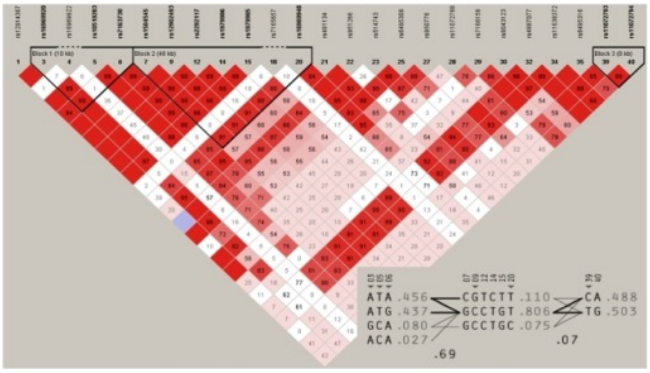
Populational linkage disequilibrium

- The chromosomes become a fine-grained mosaic of grey and black
- Called LD blocks, segments

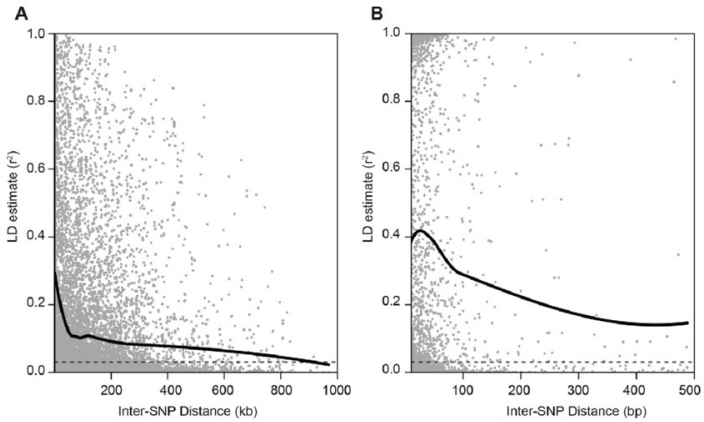


- However, complete mixture is difficult to attain
- The blocks are « fuzzy » blocks

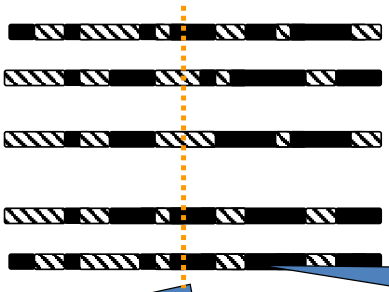
- Human


ATA .456
ATG .437
GCA .080
ACA .027
.69
CGTCTT .110
GCCTGT .806
GCCTGC .075
.07
CA .488
TG .503

- Apple


A
B
LD estimate (r²)
Inter-SNP Distance (kb)
Inter-SNP Distance (bp)

Populational linkage disequilibrium



•Some people distinguish LD and pedigree relationships
•It's pretty much the same thing

An stretch (=chromosomal segment) is conserved because it comes from the same ancestor (co-ancestry).

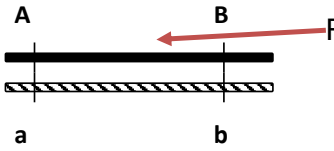
The « existence » of only a few conserved stretches at the same place creates LD.

LD is therefore:
an over-representation of segments from a few gametes
that existed in the population some time ago.

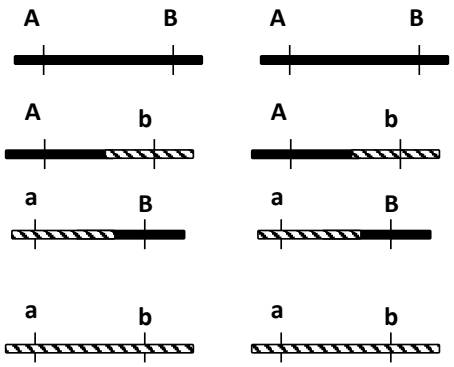
•The value of LD (e.g. r^2) observed at large distances is a function of recent relationships
•… at short distances is a function of distant relationships

Within-family linkage disequilibrium

- Consider this male who has 8 progeny



Recombination fraction: 0.50

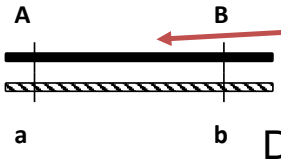


These are the chromosomes in the sons (i.e. the gametes the male transmitted)

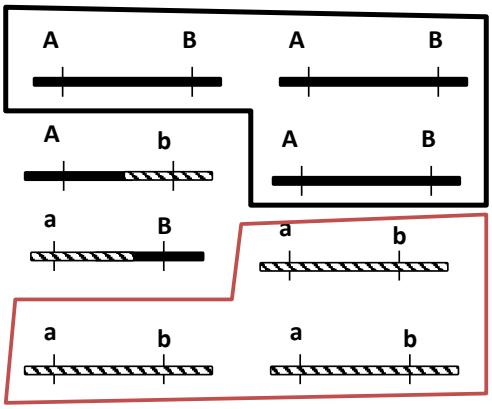
We found linkage equilibrium in one generation

Within-family linkage disequilibrium

- Consider this male who has 8 progeny



Recombination fraction: 0.25

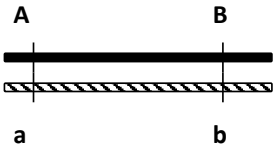


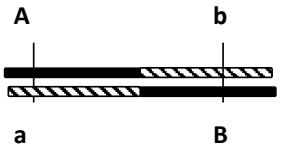
Due to non-recombination linkage disequilibrium has been generated

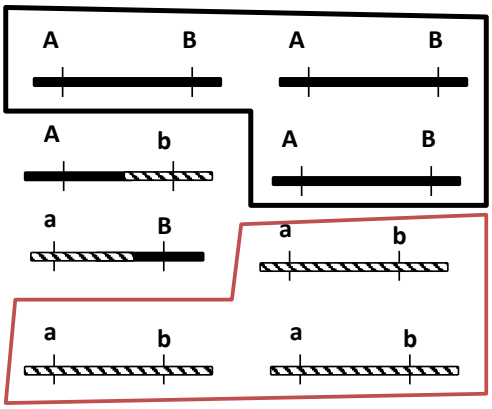
	A	a
B	0.375	0.175
b	0.175	0.375

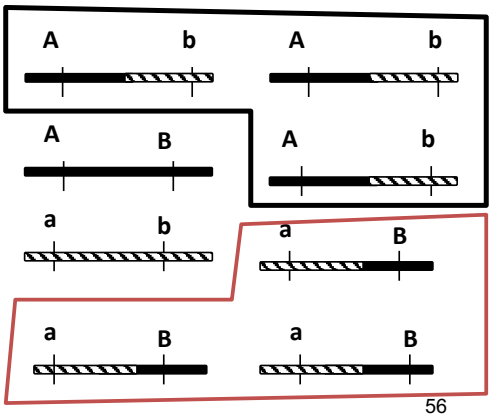
Within-family linkage disequilibrium

- Assume now there are *two* males









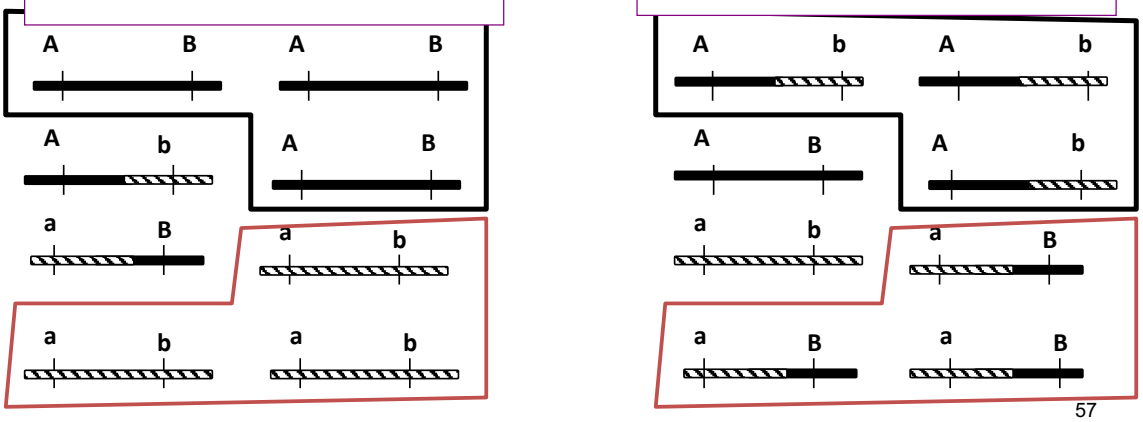
Within-family linkage disequilibrium

- Assume no

Within-family linkage disequilibrium

	A	a
B	0.375	0.175
b	0.175	0.375

	A	a
B	0.175	0.375
b	0.375	0.175

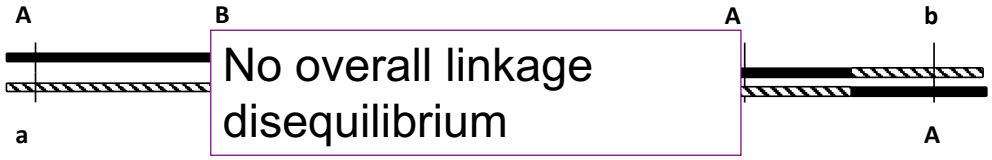


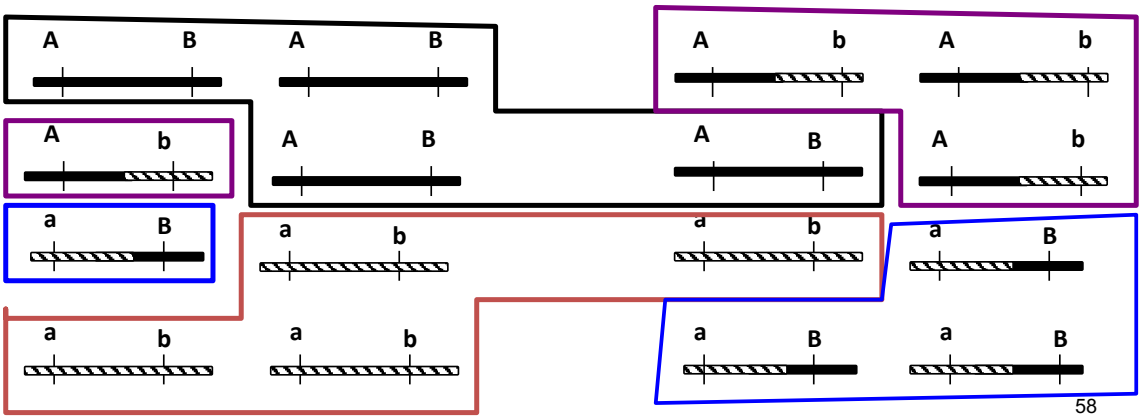
Within-family l brium

	A	a
B	0.5	0.5
b	0.5	0.5

- Assume now there

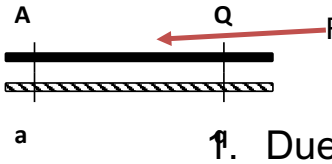
No overall linkage disequilibrium



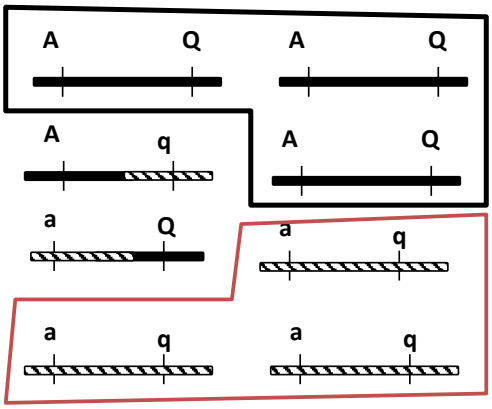


Within-family linkage analysis

- There is a QTL 25cM from A with alleles Q/q



Recombination fraction: 0.25

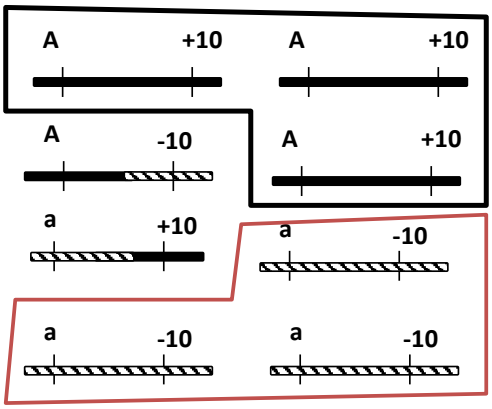


1. Due to non-recombination linkage disequilibrium has been generated

	A	a
Q	0.375	0.175
q	0.175	0.375

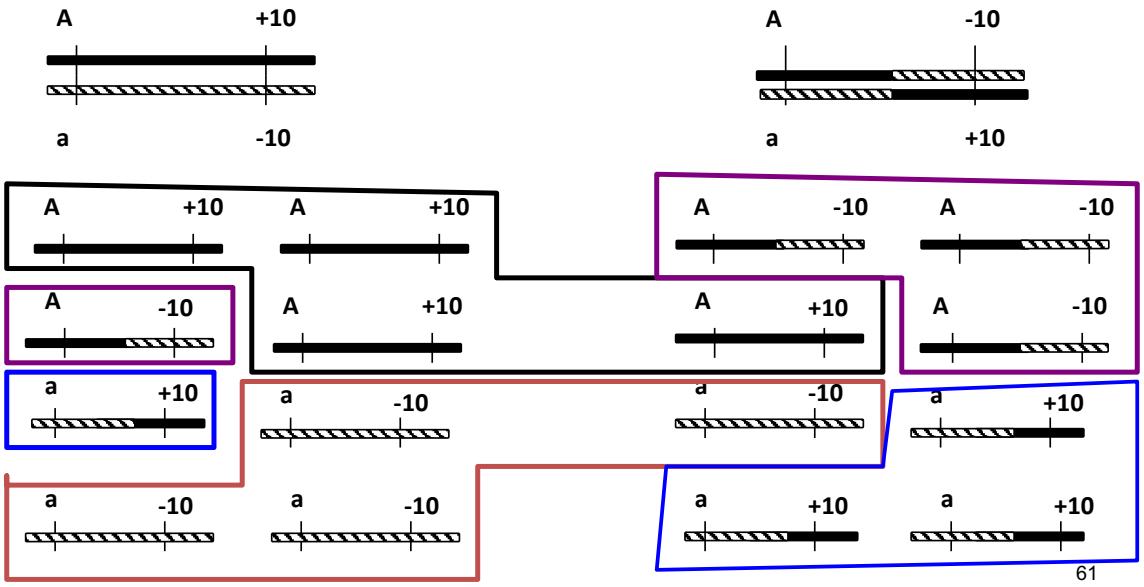
Within-family linkage analysis

- Imagine that Q=+10 kg and q=-10 kg
- The « apparent » effect of A is +5kg
- But in the other family this is reversed !



Across the two families

- Marker A has no apparent effect



Some consequences

- Markers that have “apparent” positive effect in one families may have “apparent” negative effect in other families
- These “apparent” associations break with distance
- The closest the marker and the QTL, the stronger and more stable the association
- Thus, we need many markers for associations to be stable
- Breeds are “big” families, so predicting across breeds is difficult

Measures of LD: r^2

if we use « gene content »
« A » = 1, « a »=0
« B » = 1, « b »=0
r is the correlation between two loci

$$r = \frac{f(AB) - pq}{\sqrt{p(1-p)q(1-q)}} \qquad r = \frac{D}{\sqrt{p(1-p)q(1-q)}}$$

- Not free from problems but can be understood by statisticians (and breeders)
- The sample size needed to achieve a given power is proportional to $1/r^2$ (Pritchard Przeworski 2001 Am J Hum Genet 69:1)
- Everybody uses it to describe things in genomic selection.

Properties of gene content

Data files

```
64546020 1980 6 17 15 13 4 1 2 0 0 0  1676.8800
64546020 1981 7 18 16 12 5 1 2 0 0 0  1433.6090
64546020 1982 8 17 15 14 5 1 2 0 0 0   769.2500
64546020 1980 6 17 15 12 4 1 2 0 0 0  1466.4200
64546020 1981 7 18 16 13 5 1 2 0 0 0  1474.4940
64546020 1982 8 17 16 14 5 1 2 0 0 0  1523.0290
64546020 1980 6 17 15 12 4 1 2 0 0 0  1601.2290
64546020 1981 7 17 16 12 5 1 2 0 0 0  1205.8100
64546020 1982 8 17 16 14 5 1 2 0 0 0   676.5150
64546020 1980 5 17 15 13 3 1 2 0 0 0  2122.1930
64546020 1981 6 17 15 12 4 1 2 0 0 0  2227.5940
64546020 1982 7 17 16 14 5 1 2 0 0 0  1593.4090
64546020 1980 5 17 16 11 3 1 2 0 0 0  2132.2250
64546020 1981 6 17 16 13 4 1 2 0 0 0  2100.5200
64546020 1982 7 17 16 12 5 1 2 0 0 0  1792.2250
64546020 1983 8 17 15 12 5 1 2 0 0 0  1492.0900
64546020 1984 9 17 15 11 5 1 2 0 0 0  1607.3500
64546020 1985 10 17 15 12 5 1 2 0 0 0  1534.3350
64546020 1986 11 18 17 15 6 1 2 0 0 0   958.1200
```


Pedigree files

```
00000700640031;00000000000000;00000700620012;1964;2
00000700640032;00000700620045;00000700600138;1964;2
00000700640033;00000700630065;00000700540069;1964;2
00000700640034;00000000000000;00000700580089;1964;2
00000700640035;00000000000000;00000700590106;1964;2
00000700640036;00000700630065;00000700550017;1964;2
00000700650001;00000700620047;00000700610007;1965;2
00000700650002;00000702630050;00000700560023;1965;2
00000700650003;00000700620047;00000700600125;1965;2
00000700650004;00000700620047;00000700620027;1965;2
```

Another data file

```
64000670990546 1
45214790003 1
45214790004 2
45199680012 2
45307160107 1
45199690008 2
64000249040705 1
45189980105 1
64000249030710 2
45214790006 2
45199680014 1
45214780461 2
45253180017 1
64000311010387 1
45253180018 1
45075980006 2
64000779010288 1
45315380096 2
45075980007 1
64000779010289 0
```

???

Animal

Another data file

```
64000670990546 AG
45214790003 AG
45214790004 AA
45199680012 AA
45307160107 AG
45199690008 AA
64000249040705 AG
45189980105 AG
64000249030710 AA
45214790006 AA
45199680014 AG
45214780461 AA
45253180017 AG
64000311010387 AG
45253180018 AG
45075980006 AA
64000779010288 AG
45315380096 AA
45075980007 AG
64000779010289 GG
```

Gene content !!

Animal

Gene content of marker 3

```
64000670990546 120120202102111210122210200 0
   45214790003 121111202111010210222220200 1
   45214790004 221110201101021200222211211 0
   45199680012 210111102011011210122201211 0
   45307160107 121210202002022200222211211 0
   45199690008 220210202001022211222210211 1
64000249040705 121200202001022210122201211 0
   45189980105 120110202101111220022200200 0
64000249030710 221110201101112211222201222 1
   45214790006 221200202000021211122210101 1
   45199680014 122201101101012222222111111 0
   45214780461 221210201102022201122221111 1
   45253180017 110111102000202221222210222 2
64000311010387 122200202001021210122201211 0
```


Gene content

- Gene content (GC) is the number of copies of the reference allele (for instance "A")
- We call gene content *z* in this notes and slides
- $z = \begin{cases} 0 \\ 1 \\ 2 \end{cases}$ for genotypes $\begin{cases} \text{"}no\ reference\ allele\text{"} \\ 1\ \text{"}reference\ allele\text{"} \\ 2\ \text{"}reference\ allele\text{"} \end{cases}$
- What properties does gene content have, as a "trait"?

Gene content mean and variance

- $\bar{z} = 2p$
- $\sigma_z^2 = E(z^2) - E(z)^2 = 2pq$ if there is HWE

Heritability of gene content

- If the genotype is accurate, the trait *z* is observed with no error
- *z* is transmitted from parents to offspring and there is no external influences
- *z* is additive (by definition)
- Heritability of *z* is 1 (!!!)

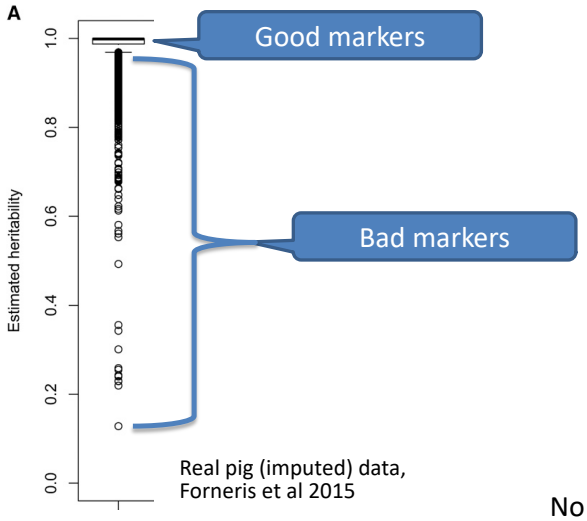
We can model gene content as a quantitative trait:

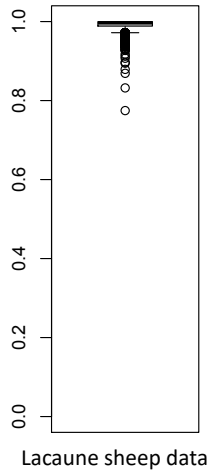
- $Cov(z_i, z_j) = A_{ij} 2pq$ (Cockerham - explain)
- $\boldsymbol{z} = \mathbf{1}\mu + \boldsymbol{u} = \mathbf{1}(2p) + \boldsymbol{u}$
- $\mathrm{V}ar(u) = \boldsymbol{A}\sigma_u^2 = \boldsymbol{A}\sigma_z^2 = \boldsymbol{A}2pq$

Gene content as a quantitative trait

- We can estimate heritability of gene content
  - Extract one marker from the genotype file and treat it as data
  - Estimate heritability by REML
  - It should give $\hat{h}^2 \approx 0.99$ or similar
  - p-value of $\hat{h}^2 \neq 1$ using LRT

Quality control using heritability of gene content





8% rejected markers (p<0.01)
Why do we have bad markers?
Probably due to poor imputation

No rejected markers based on LRT
Why do we have good markers?
Good pedigree recording, DNA sampling, and imputation

Quality control using heritability of gene content

- No one checks h^2 of gene content by default, but it is very useful to detect horrible mistakes
- In small data sets (<5000 animals with genotypes), it takes minutes in *preGSf90*
- *qcf90* does it in a few hours for large data sets

Imputation

- What do we mean by imputation?
- "Guess" the missing marker
- Why is this useful

  (1) For software that don't admit missing values at genotypes: fill-in the small holes like

  – 0121102212115010055111 12000

Missing !!

Imputation

(2) to use "cheap chips"

Missing !!

- We may have big holes
  - $0555_2555551555551555_0$
  - Low density chips: impute from 6K to 50K
  - Very high density chips: impute from 50K to 700K
  - Very low density chips: impute from 1000 to 50K

Crude imputations

<u>Not</u> recommended

- Draw genotype from HW distribution:
  - {AA,AG,GG} with probabilities $\{p^2, 2pq, q^2\}$
  - Will lead to parent/offspring incompatibilities
- Use heterozygotes
  - Will lead to too many heterozygotes

Strategies for imputation

- Family based

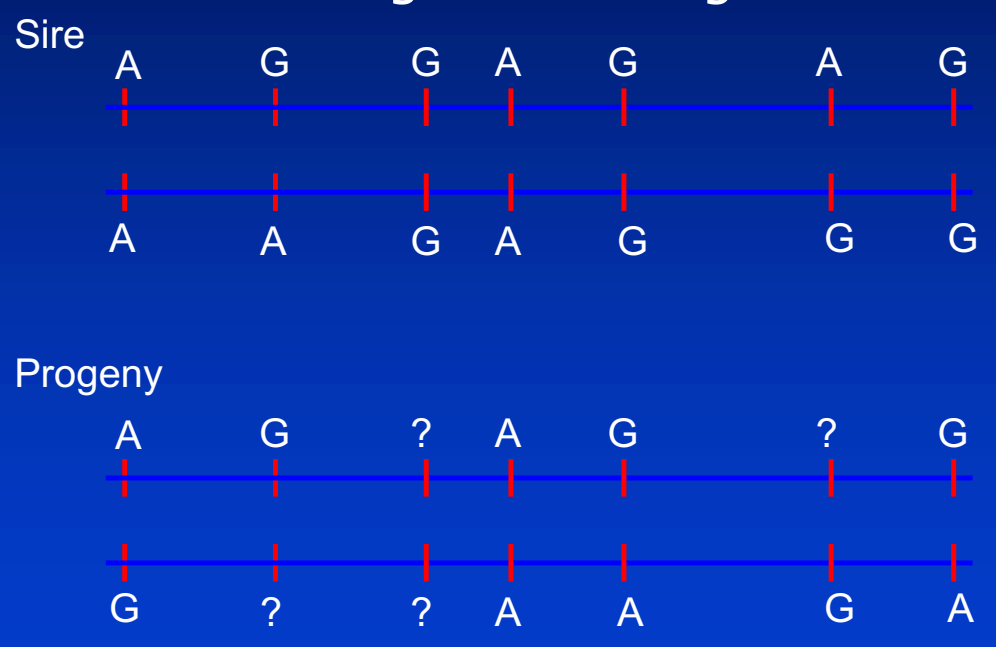
  We compare chromosome chunks transmitted from parents to offspring and fill-in the holes

- Population based

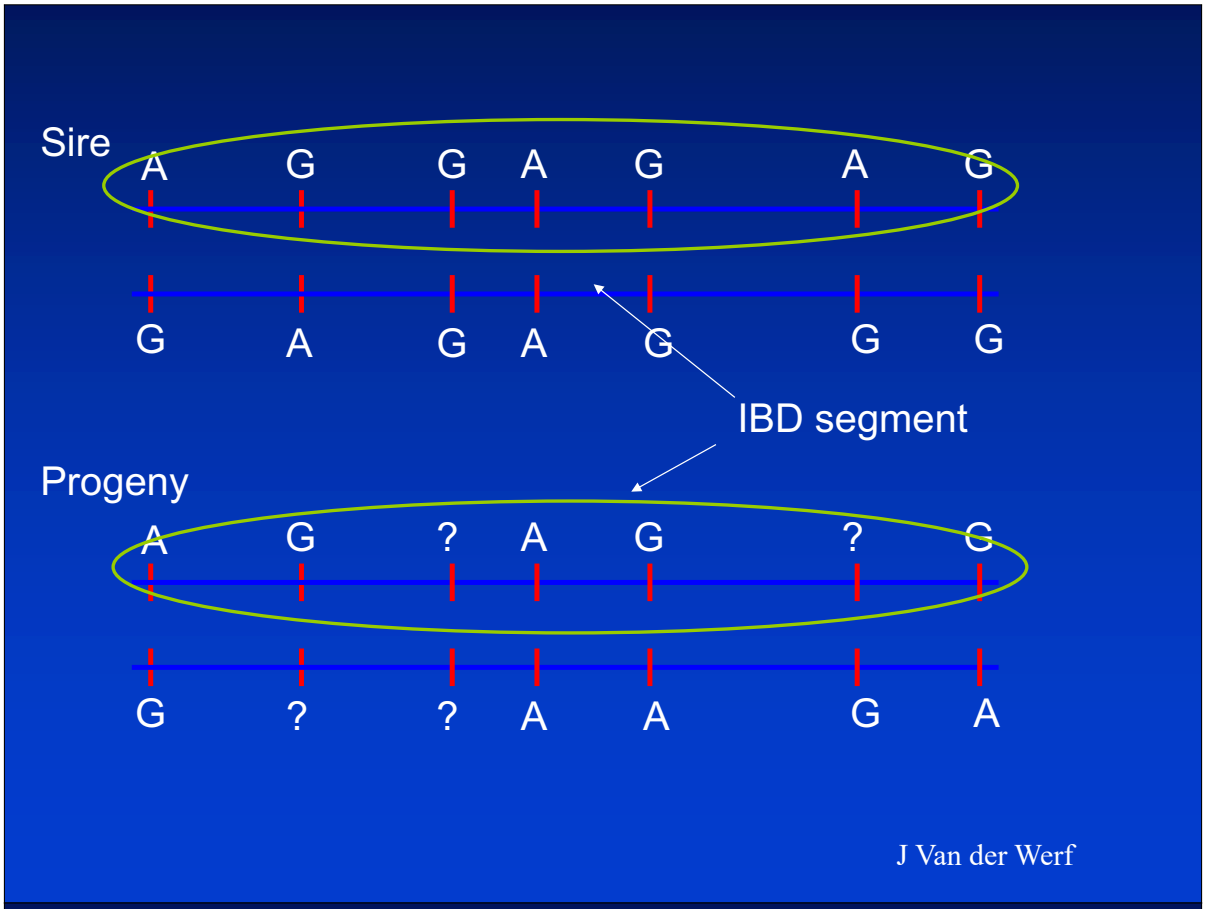
  We (roughly) make a library of existing haplotypes and compare to our incomplete haplotype

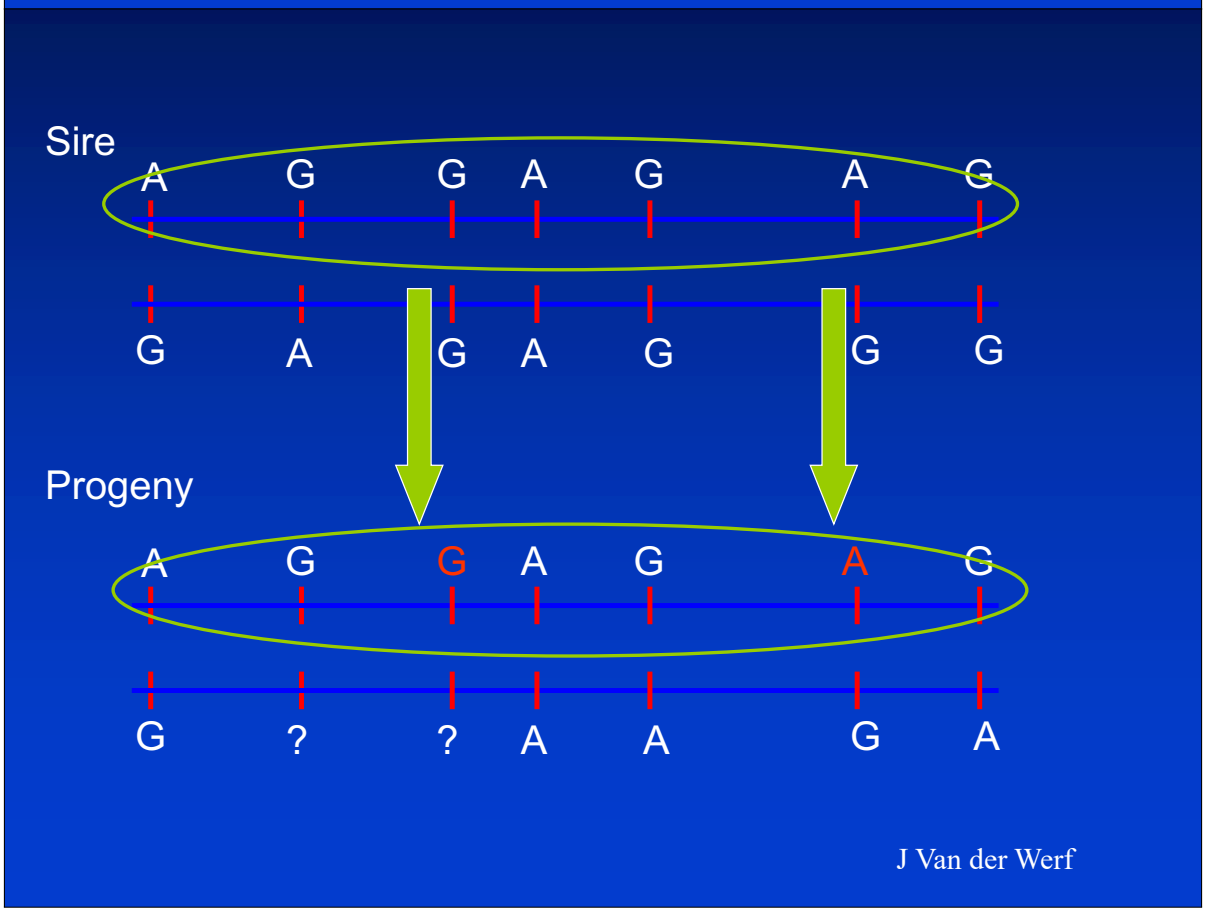
- Imputation is based on looking at neighboring markers
- We need a map of the genome !!

Finding an IBD segment

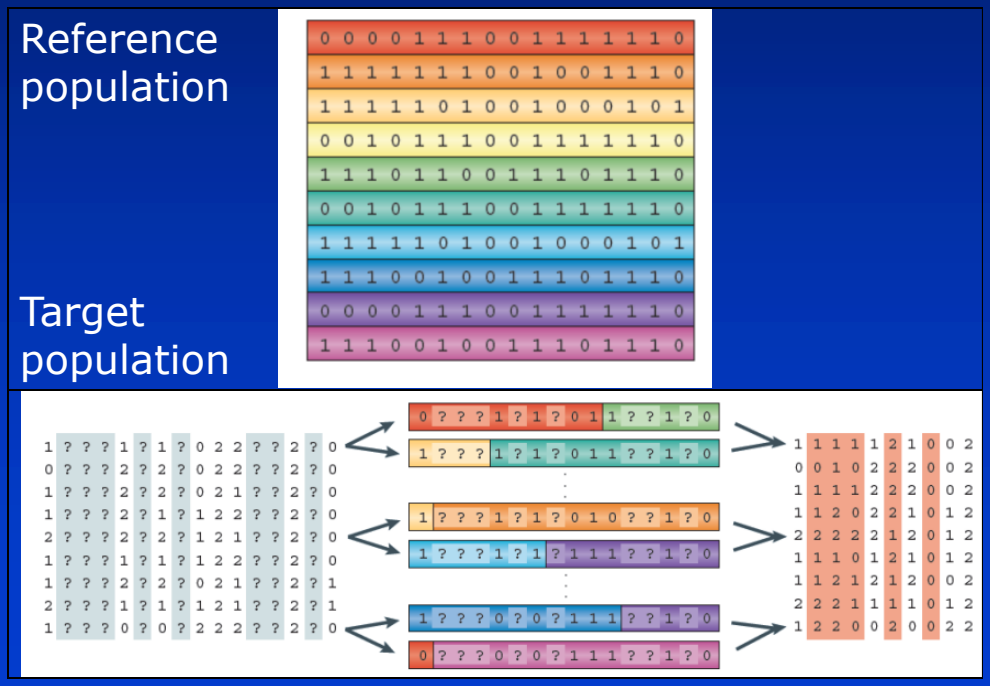


J Van der Werf


Sire
A G G A G A G
G A G A G G G
IBD segment
Progeny
A G ? A G ? G
G ? ? A A G A
J Van der Werf


Sire
A G G A G A G
G A G A G G G
Progeny
A G G A G A G
G ? ? A A G A
J Van der Werf

Population based imputation



Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010 11:499-511.

Typical outputs from imputation

- Accuracy = correlation of real and imputed genotypes
- Concordance = percentage (%) of genotypes called correctly
- Concordance is a bad metrics because genotypes will be imputed correctly just by chance

Typical pitfalls from imputation

- Several horror stories linked to imputation
  - Very small SNP chips (<6K) typically impute very poorly
  - Pedigrees and DNA sample identification need to be perfect
  - To impute correctly, the reference population (a set of individuals fully genotyped at >50K) needs to cover the entire genetic variation. I can't impute Scottish Angus from Angus.

- Errors in imputation may go undetected, but then they create contradictory informations for ssGBLUP
- Imputation tends to create too many heterozygotes
- LD chips + imputation is not a substitute for 50K genotyping
- You better test what you're doing

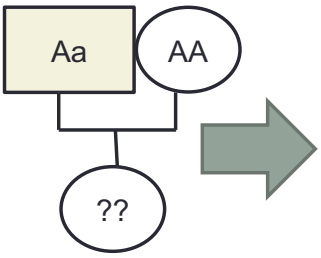
Non genotyped animals

- If animals do not have any genotype for any marker, what can we do?
- A few of them can be "imputed" classically
  - if they have large offsprings genotyped, e.g. 5 offspring for an embryo transfer dam
- In most cases this is impossible
- We still can use "linear" imputation

Linear imputation

- Gengler et al. (2007) conceived an algebraic way to obtain regression-based point estimates of genotypes (== to McPeek et al. 2004)
- Christensen & Lund (2010) showed how to take the variation into account

- Genotype of descendants = half their parents + Mendelian sampling



AA with probability ½
Aa with probability ½

E(Genotype) = $\frac{3}{2}$"A" + $\frac{1}{2}$"a"

Extending to all the pedigree…

Genotype prediction using BLUP for gene content (Gengler's method)

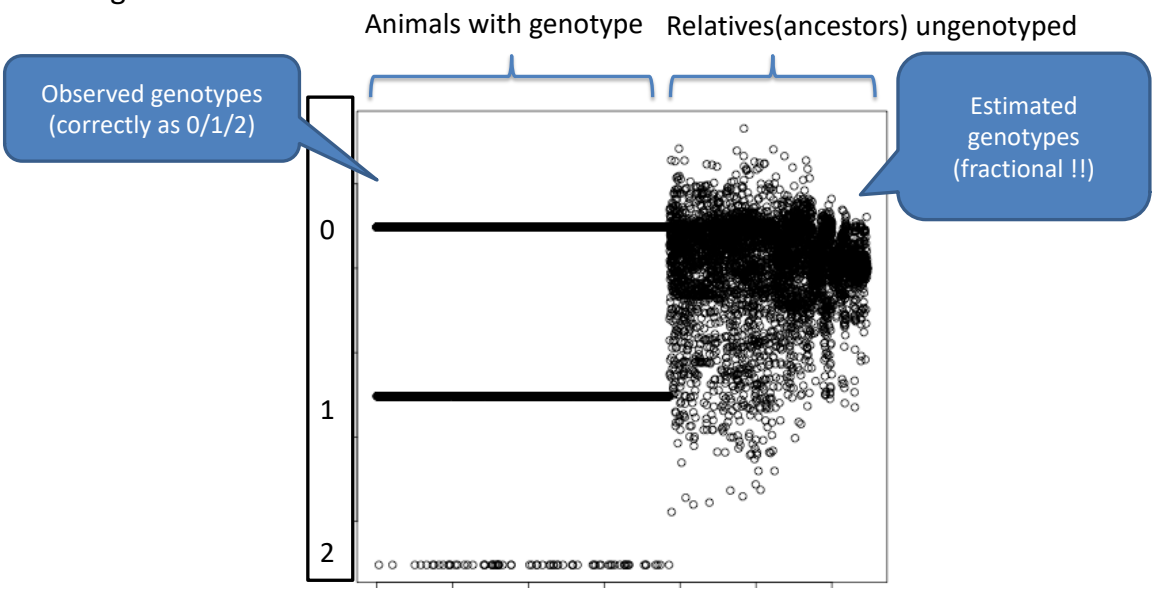
- Assuming $h^2 \approx 0.99$, use BLUP !!
- $\boldsymbol{z} = \mathbf{1}\mu + \boldsymbol{W}\boldsymbol{u} + \boldsymbol{e}$

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{W} \\ \boldsymbol{W}'\boldsymbol{Z} & \boldsymbol{W}'\boldsymbol{W} + \boldsymbol{A}^{-1}\boldsymbol{\lambda} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\mu}} \\ \widehat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{z} \\ \boldsymbol{W}'\boldsymbol{z} \end{pmatrix}$$

- On exit, $\mathbf{1}\hat{\mu} + \widehat{\boldsymbol{u}}$ are estimates of gene content for all animals
- And $\frac{\hat{\mu}}{2}$ is an estimate of p in the base generation

Example

- Pig data



Why is linear imputation bad?

- It is very little accurate
- Far animals tend all to be identical to $2\hat{p}$
- Uncertainty in linear imputation is ignored

- But it sets the stage for SSGBLUP

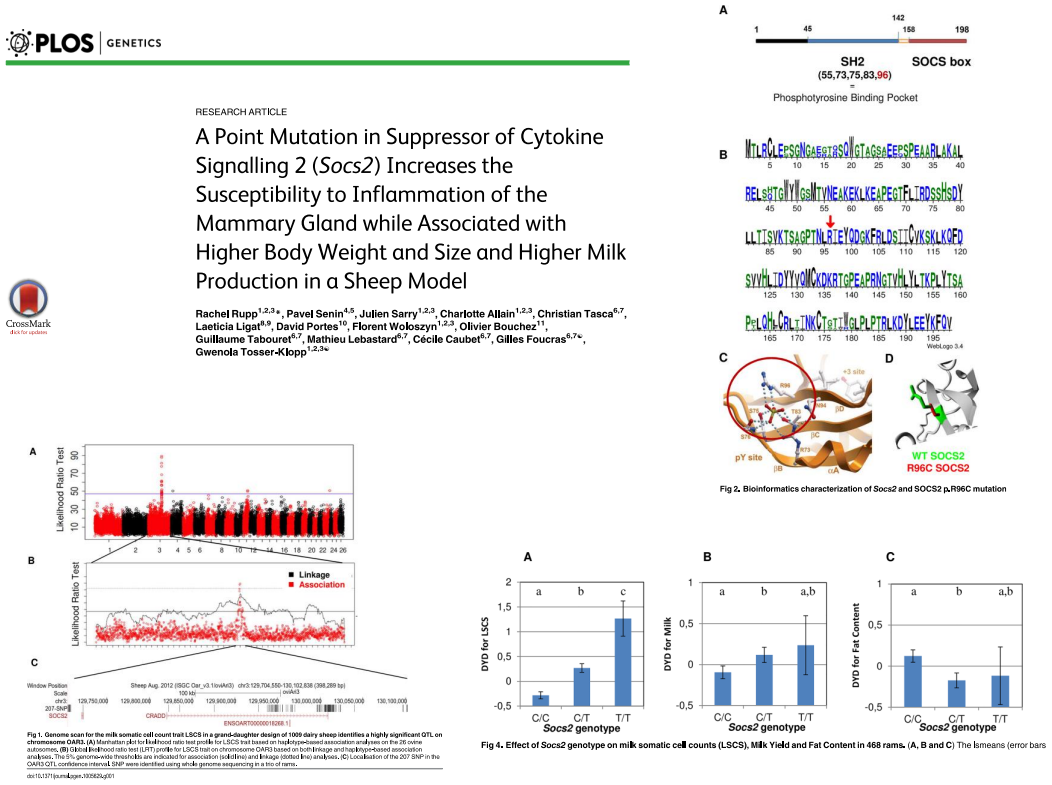
Marker-based models for Genomic selection

- Single QTL
- Whole-genome (multiple marker) genomic selection

Single QTL

Assume that we know a large effect QTL (a major gene)

- the halothane gene (HAL)
- the α_{s-1} caseine in dairy goats
- DGAT1
- SOCS2
- BMP15
- IFG-2
- GHR



PLOS | GENETICS

RESEARCH ARTICLE

A Point Mutation in Suppressor of Cytokine Signalling 2 (*Socs2*) Increases the Susceptibility to Inflammation of the Mammary Gland while Associated with Higher Body Weight and Size and Higher Milk Production in a Sheep Model

Rachel Rupp[1,2,3]*, Pavel Senin[4,5], Julien Sarry[1,2,3], Charlotte Allain[1,2,3], Christian Tasca[6,7], Laeticia Ligat[8,9], David Portes[10], Florent Woloszyn[1,2,3], Olivier Bouchez[11], Guillaume Tabouret[6,7], Mathieu Lebastard[6,7], Cécile Caubet[6,7], Gilles Foucras[6,7]*, Gwenola Tosser-Klopp[1,2,3]*

Fig 2. Bioinformatics characterization of *Socs2* and SOCS2 p.R96C mutation

Fig 1. Genome scan for the milk somatic cell count trait LSCS in a grand-daughter design of 1009 dairy sheep identifies a highly significant QTL on chromosome OAR3. (A) Manhattan plot for likelihood ratio test profile for LSCS trait based on haplotype-based association analyses on the 26 ovine autosomes, (B) Global likelihood ratio test (LRT) profile for LSCS trait on chromosome OAR3 based on both linkage and haplotype-based association analyses. The 5% genome-wide thresholds are indicated for association (solid line) and linkage (dotted line) analyses. (C) Localisation of the 207 SNP in the OAR3 QTL confidence interval. SNP were identified using whole genome sequencing in a trio of rams.

Fig 4. Effect of *Socs2* genotype on milk somatic cell counts (LSCS), Milk Yield and Fat Content in 468 rams. (A, B and C) The lsmeans (error bars

3

Single QTL

Put the QTL as a fixed effect and estimate it

- $\mathbf{y}_i$= QTL effect in animal *i + e*

We can include an additional polygenic genetic value of animal *i*

- $\mathbf{y}_i$= polygenic effect of animal *i* + QTL effect in animal *i + e*

How do we do this in practice? Using linear regression

Multiallelic QTL

assume that we have a

four-allele $\{A, B, C, D\}$ locus

and three individuals with genotypes

$$\{BC, AA, BD\}$$

Base model

- Genotypes are $\{BC, AA, BD\}$
- Data is $\mathbf{y} = \begin{pmatrix} 12 \\ 35 \\ 6 \end{pmatrix}$

3 individuals, 1 marker with 4 alleles

$$\boldsymbol{Za} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_A \\ a_B \\ a_C \\ a_D \end{pmatrix}$$

Single QTL regression

this gives

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

$$\begin{pmatrix} 12 \\ 35 \\ 6 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \mu + \begin{pmatrix} 0 & 1 & 1 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_A \\ a_B \\ a_C \\ a_D \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}$$

- Can be solved by least squares

Single QTL regression with polygenic based on pedigree

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Za} + \boldsymbol{Wu} + \boldsymbol{e}$$

$$\begin{pmatrix} 12 \\ 35 \\ 6 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \mu + \begin{pmatrix} 0 & 1 & 1 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_A \\ a_B \\ a_C \\ a_D \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}$$

- $Var(\boldsymbol{u}) = \boldsymbol{A}\sigma_u^2$
- Can be solved by BLUP

Goddard, M. E. (2003). Animal breeding in the (post-) genomic era. *Animal Science*, *76*(3), 353-365.

1. Although it is possible to use genetic markers linked to genes of economic importance, tests for the genes themselves will be much more successful.
2. Finding these genes, that have relatively small effects, is more difficult than finding genes for a classical Mendellian trait but, as the genomic tools become more powerful, it is becoming feasible and some successes have already occurred
3. Tools such as genomic sequence, EST collections and comparative maps make this approach feasible. Candidate genes can be selected based on functional data such as gene expression
4. in the future, with many QTL identified and inexpensive genotyping combined with decreased generation intervals, large gains are possible.

- Wait, we still don't know where genes are?
- Don't we use GWAS to find them?
  - GWAS is too complicated and can find just a few genes
  - in the Notes you have long explanations
  - the fact is, most causal genes for most traits for most species are just unknown
- Meuwissen et al. 2001 proposed to use marker effects directly

Basic principle of genomic selection

- every marker has an effect on the character (even if it doesn't look like it !).
- Markers are not QTLs but
- when there are many markers,
- for each animal, "the sum of the effects of the markers" is a good predictor of "the sum of the effects of the QTLs".
- you can be a good predictor without being « real » (e.g. herd is a proxy for farmer)

Basic principle of genomic selection

- Suppose the true model is.
- $u = \sum z_i^Q \hat{a}_i^Q$, sum of effects in the QTL.
- We use an approximate model
- $u \approx \sum z_i^M \hat{a}_i^M$, "sum of effects in the markers."

- It works (although nobody quite understands how) it was a gamble ☺ and it worked.

- Other models (linkage, haplotypes,...) can be thought and used, but the model with markers is simple and analytically and computationally very grateful.

n marker regression

We estimate the effect of markers by regression

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Z_1 a_1} + \boldsymbol{Z_2 a_2} + \cdots \boldsymbol{e}$$

2-locus multiallelic marker additive model

three individuals with genotypes

$$\{BC\ EE$$
$$, AA\ EF$$
$$, BD\ FF\}$$

$$\boldsymbol{Za} = \begin{pmatrix} 0 & 1 & 1 & 0 & \vdots & 2 & 0 \\ 2 & 0 & 0 & 0 & \vdots & 1 & 1 \\ 0 & 1 & 0 & 1 & \vdots & 0 & 2 \end{pmatrix} \begin{pmatrix} a_A \\ a_B \\ a_C \\ a_D \\ \cdots \\ a_E \\ a_F \end{pmatrix}$$

But SNPs are biallelic

2-locus biallelic marker additive model

three individuals with genotypes

$$\{BA\ EE, AA\ EF, BB\ FF\}$$

This looks redundant

$$\boldsymbol{Za} = \begin{pmatrix} 1 & 1 & \vdots & 2 & 0 \\ 2 & 0 & \vdots & 1 & 1 \\ 0 & 2 & \vdots & 0 & 2 \end{pmatrix} \begin{pmatrix} a_A \\ a_B \\ \cdots \\ a_E \\ a_F \end{pmatrix}$$

if we reduce the effects to one effect per marker, we get

$$\boldsymbol{Za} = \begin{pmatrix} 1 & \vdots & 2 \\ 2 & \vdots & 1 \\ 0 & \vdots & 0 \end{pmatrix} \begin{pmatrix} a_A \\ \cdots \\ a_E \end{pmatrix} \text{ but could be } \boldsymbol{Za} = \begin{pmatrix} 1 & \vdots & 0 \\ 0 & \vdots & 1 \\ 2 & \vdots & 2 \end{pmatrix} \begin{pmatrix} a_B \\ \cdots \\ a_F \end{pmatrix}$$

4-locus biallelic marker additive model

three individuals with genotypes

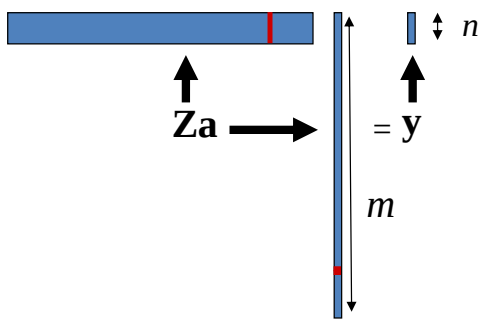
$$\{BA\ EE\ HG\ OP, AA\ EF\ GG\ OO, BB\ FF\ HH\ PP\}$$

$$\boldsymbol{Za} = \begin{pmatrix} 1 & \vdots & 2 & \vdots & 1 & \vdots & 1 \\ 2 & \vdots & 1 & \vdots & 2 & \vdots & 2 \\ 0 & \vdots & 0 & \vdots & 0 & \vdots & 0 \end{pmatrix} \begin{pmatrix} a_A \\ \cdots \\ a_E \\ \cdots \\ a_G \\ \cdots \\ a_P \end{pmatrix}$$

20-loci

$$\begin{bmatrix} 1 & 1 & 2 & 2 & 1 & 0 & 0 & 1 & 0 & 0 & 2 & 0 & 0 & 2 & 0 & 2 & 2 & 0 & 1 & 1 \\ 0 & 1 & 2 & 1 & 2 & 1 & 0 & 1 & 2 & 2 & 2 & 2 & 0 & 2 & 1 & 0 & 1 & 0 & 0 & 1 \\ 2 & 0 & 2 & 0 & 0 & 2 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 2 & 2 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

50000 loci



As many loci as you want

Fortunately we have matrix algebra

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{a} \ldots$$

- $\boldsymbol{Z}$: as many columns as markers
- $\boldsymbol{Z}$: as many rows as individuals

Prediction equations

What's all this about?

- I want to select the best dairy sheep rams, at their birth
- Predict breeding values based on a "reference population" with data and…
  - Pre-genomic: pedigree
  - Genomic: markers

Idealized process of pedigree prediction

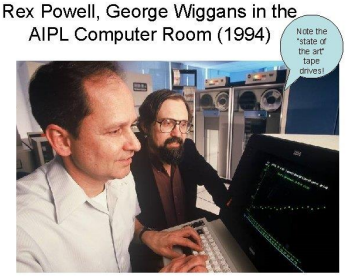
In the reference population:

Get pedigree ($\boldsymbol{A}$)
Get phenotypes ($\boldsymbol{y}$)

Estimate Breeding Values $\boldsymbol{u}$ in the reference population from

$$\boldsymbol{y} = \boldsymbol{1}\mu + \boldsymbol{u_r} + \boldsymbol{e}\,,$$

Rex Powell, George Wiggans in the AIPL Computer Room (1994)



In the candidates: estimate Breeding Values from relationships in $\boldsymbol{A}$ and estimates in the reference:

$$\hat{u}_c = \boldsymbol{A}_{cr}\boldsymbol{A}_{rr}^{-1}\hat{\boldsymbol{u}}_{\boldsymbol{r}}$$

or (for progeny with no other data)

$$\hat{u}_c = \hat{u}_d/2 + \hat{u}_s/2$$

Idealized process of genomic prediction

In the reference population:

Get markers' genotypes ($\boldsymbol{Z}_r$)
Get phenotypes ($\boldsymbol{y}$)

Estimate markers effects $\boldsymbol{a}$ from

$$\boldsymbol{y} = \boldsymbol{1}\mu + \boldsymbol{Z}_r\boldsymbol{a} + \boldsymbol{e}\,,$$

In the candidates:

Get markers' genotypes ($\boldsymbol{Z}_c$)
Take estimates $\hat{\boldsymbol{a}}$ from above
Estimate breeding values as

$$\hat{\boldsymbol{u}}_c = \boldsymbol{Z}_c\hat{\boldsymbol{a}}$$

New animal

- I know from the reference population that SNP effects are estimated as $\hat{\boldsymbol{a}} = \begin{pmatrix} 0.1 \\ 1.1 \\ -2 \end{pmatrix}$
- I genotype the animal and is $\begin{pmatrix} AA \\ GC \\ AG \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}$ 
- Its breeding value is $\hat{\mathrm{u}} = \begin{pmatrix} 2 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0.1 \\ 1.1 \\ -2 \end{pmatrix} = -0.7$

From marker effects to breeding values

- Once we have estimates of marker effects, $\hat{\boldsymbol{a}}$
- For any animal (young or old, with or without data) the GEBV is $\hat{\boldsymbol{u}} = \sum z_i \hat{a}_i = \boldsymbol{z}\hat{\boldsymbol{a}}$
- Note that **Z** must always be encoded in the same way....
- (Why is this a GEBV)?

From marker effects to breeding values

- At one locus, a ram has a certain genotype, say GT, which is coded as z (e.g. $z = 1$)
- It is going to pass to its offspring, ½ of the times "T" and ½ of the times "G", so the offspring will receive on average $\frac{z}{2}$
- If the a effect of the marker is conserved in the progeny, then *on average* the offspring will have $\frac{z}{2}a$ from the ram, so the ram's EPD will be $\frac{z}{2}\hat{a}$ and its EBV=2*EPD will be $z\hat{a}$.
- That is, using the "additive" coding {0,1,2} ($\pm$ a constant) of the genotypes leads naturally to obtain (G)EBVs.
- This is not a property of other "relationships", (*e.g.* kernel matrices with Euclidean distances)

- How do we estimate marker effects?
- By the time-honored technique of Regression

Least Squares estimate of marker effects

$$\boldsymbol{Z'Z\hat{a} = Z'y}$$

```
do i=1,nanim
        read(1,'(a14,1x,60000i1)') y(i),Z(i,:)
enddo
ZpZ=matmul(transpose(Z),Z)
Zpy=matmul(transpose(Z),y)
ZpZ=ginv(ZpZ)
a=matmul(ZpZ,Zy)
end
```

Read Z and y

Build Z'Z and Z'y

Solve $\hat{a} = (Z'Z)^{-1}Z'y$

Really?

Estimating SNP effects

- The simultaneous estimates of many markers by least squares are very poor, if we have more SNPs than individuals
- Even if we had many individuals, there is a missing piece of information:
  - most SNPs should *not* have a large effect
  - this is a « prior » information
- Can we do something?
- We should use the theory of « Best Prediction » or « Bayesian Regressions »

Bayesian regressions

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{a} + \cdots + \boldsymbol{e}$$

- Everyone assumes $p(\boldsymbol{e}) \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{R})$
- what do we assume for marker effects: $p(\boldsymbol{a})$
- Do we want very strong marker effects?
  - No: $p(\boldsymbol{a}) \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I}\sigma_a^2)$ SNP-BLUP == Ridge Regression == rrBLUP
  - Yes: Bayesian Alphabet (Bayes A, B, C, R, S… Bayesian Lasso…)
    - see Notes for all these methods
    - usually they don't improve predictions
    - "effect of prior vanishes with more data"

- Effect sizes are misleading
- It is quite difficult to know if genes are there
- Markers around capture the effect of the gene anyway

why methods (don't) matter

Should we use the single nucleotide polymorphism linked to *DMRT3* in genomic evaluation of French trotter?[1]

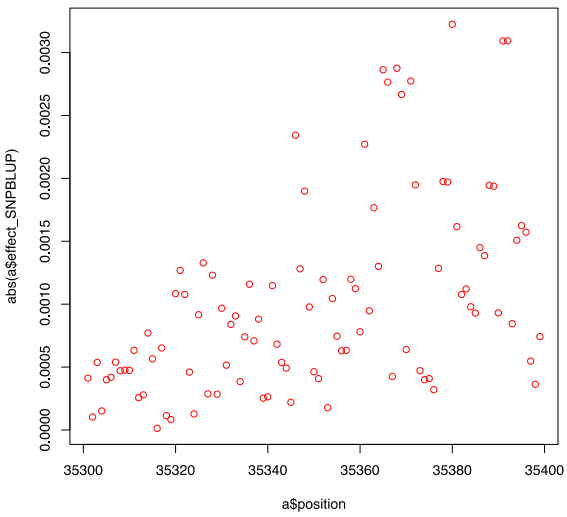
S. Brard*†‡[2] and A. Ricard§#

Does heterozygosity at the *DMRT3* gene make French trotters better racers?

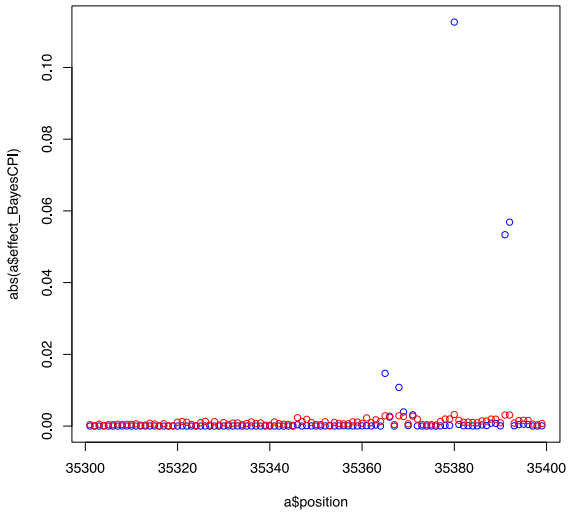
Abstract

Background: Recently, a mutation was discovered in the *DMRT3* gene that controls pacing in horses. The mutant allele A is fixed in the American Standardbred trotter breed, while in the French trotter breed, the frequency of the wild-type allele C is still 24%. This study aimed at measuring the effect of *DMRT3* genotypes on the performance of French trotters and explaining why the polymorphism still occurs in this breed. Using a mixed animal model,

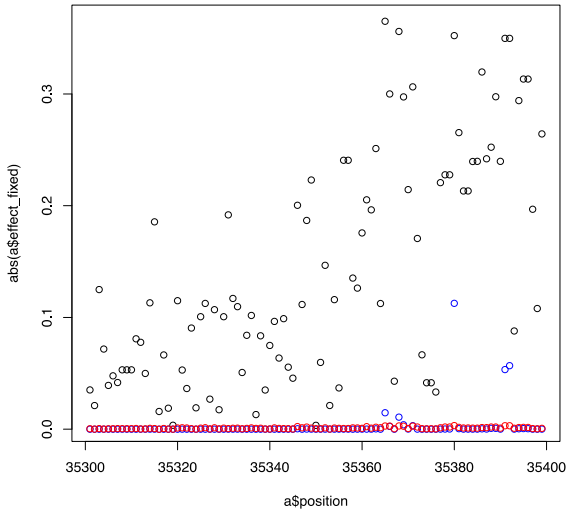
Effect estimated by SNP-BLUP



Effects estimated by BayesCPi

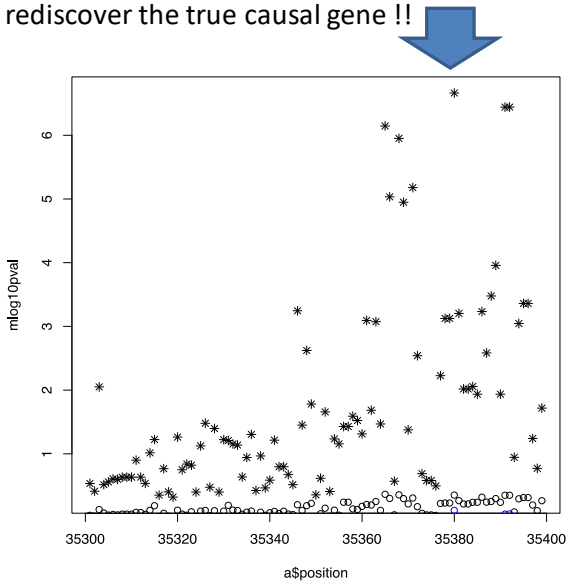


Effect estimated by separate Single marker regressions



-log10(P-values) from the separate Single marker regressions

Finally we rediscover the true causal gene !!



Remember the SOCS2 gene?

Alternative methods improve the accuracy of genomic prediction using information from a causal point mutation in a dairy sheep model

Claire Oget[1*], Marc Teissier[1], Jean-Michel Astruc[2], Gwenola Tosser-Klopp[1†] and Rachel Rupp[1†]

- "In the absence of chip data, [including the major gene as a covariate], compared to pedigree-based BLUP, efficiently accounted for [...] genotyping information on SOCS2 as accuracy was increased by 6.25%"
- "Adding the SOCS2 SNP to ssGBLUP methods led to an average gain of 0.26%."

- In fact, SOCS2 strong effect is well captured by neighboring SNPs – even in ssGBLUP (which is like SNP-BLUP)
- fitting SOCS2 explicitly to "extract" its large effect does not improve anything

SNP-BLUP

- After 10 years of experimentation, normality of marker effects is a good assumption
- This assumption of normality is called in different contexts
  - BLUP
  - genomic BLUP
  - SNP-BLUP
  - GBLUP
  - ridge regression
  - Random RegressionBLUP
- I will keep GBLUP for the use of the genomic relationship matrix
- and SNP-BLUP for the direct estimation of SNP effects

Mixed model equations for SNP-BLUP

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

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \sigma_a^2 = \mathbf{I}\sigma_a^2$$

- **Z'Z** is *not* diagonal
- Prior information: variance of SNP effects
- usually assumed $Var(\boldsymbol{a}) = \boldsymbol{D} = \boldsymbol{I}\sigma_a^2$

SNP-BLUP is flexible

- In theory
  - Multiple trait models
  - REML
  - Threshold models
  - Maternal effects, random regression, social effects…
- But:
  - Little software around
  - Multiple trait models will involve huge matrices

Coding

Coding: How do we fill **Z** based on genotypes

- This is a frequent source of confusion even for experienced people
- It is mixed with shifting the mean and variance of EBV
- The main message is that "it does not matter" if you are coherent through all the steps in your research
  - (for SNP-BLUP and GBLUP; not for ssGBLUP)
- The notes (should) contain all the gory details
- most details are in Strandén & Christensen (GSE 2011)

Coding

- Reference allele -> sign of marker effects
- “centering” -> shift of the overall mean
- “scaling” -> shift of the implicit genetic variance

- Assume that we use SNP-BLUP equations
- Importantly, we keep σ_{a0}^2 fixed across the different codings

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{I}\sigma_{a0}^{-2} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}$$

let’s check this now

Coding: reference allele

three individuals with genotypes

$\{BA\ EE$
$, AA\ EF$
$, BB\ FF\}$ ⇒ $\boldsymbol{Za} = \begin{pmatrix} 1 & \vdots & 2 \\ 2 & \vdots & 1 \\ 0 & \vdots & 0 \end{pmatrix} \begin{pmatrix} a_A \\ \cdots \\ a_E \end{pmatrix}$

but could equally be

$$\boldsymbol{Z}^*\boldsymbol{a}^* = \begin{pmatrix} 1 & \vdots & 0 \\ 0 & \vdots & 1 \\ 2 & \vdots & 2 \end{pmatrix} \begin{pmatrix} a_B \\ \cdots \\ a_F \end{pmatrix}$$

This depends on the "<u>reference allele</u>". It doesn't matter which one we take

If the other allele is used as reference, then the numbers in $\boldsymbol{Z}$ are reversed. In fact $\boldsymbol{Z}^* = 2\mathbf{1}\mathbf{1}' - \mathbf{Z}$, as a result $\widehat{\boldsymbol{a}^*} = -\widehat{\boldsymbol{a}}$ (because of properties of [Mixed Models, Bayesian] regression models)

Hence, $\widehat{\boldsymbol{u}}^* = \boldsymbol{Z}^*\widehat{\boldsymbol{a}^*} = (2\mathbf{1}\mathbf{1}' - \boldsymbol{Z})\widehat{-\boldsymbol{a}} = \mathbf{1}\mu + \boldsymbol{Z}\widehat{\boldsymbol{a}} = \boldsymbol{Z}\widehat{\boldsymbol{a}} = \widehat{\boldsymbol{u}}$
with $\mu = 2\sum \hat{a}_i = 0$ (because of properties of [Mixed Models, Bayesian] regression models)

Coding: reference allele: New animal

- Do NOT make the mistake of coding in the opposite (or just a different) way new batches of animals
- In the reference population SNP effects were estimated as $\widehat{\boldsymbol{a}} = \begin{pmatrix} 0.1 \\ -2 \end{pmatrix}$ and the reference allele were {A,E}
- I genotype a new animal and is $(BA \quad FF) \Longrightarrow \boldsymbol{z} = (1 \quad 0)$
  - Its breeding value is $\hat{u} = (1 \quad 0)\begin{pmatrix} 0.1 \\ -2 \end{pmatrix} = 0.1$
- if we (wrongly) take reference alleles {B,F} then
  - $\hat{u} = (1 \quad 2)\begin{pmatrix} 0.1 \\ -2 \end{pmatrix} = 3.9$ …wrong !!

Coding: « centering »

Genotype	101 Coding	012 Coding	Centered coding
aa	$-a_i$	0	$-2p_i a_i$
Aa	0	a_i	$(1-2p_i)a_i$
AA	a_i	$2a_i$	$(2-2p_i)a_i$

In all cases $\boldsymbol{Z}^* = \mathbf{Z_{012}} - \mathbf{2p}^{*\prime}$ where $\mathbf{p}^*$ has all possibilities:

- 0 (« 012 Coding »)
- 0.5 (« 101 Coding ») ,
- observed allele frequencies (« Centered coding »)
- base population allele frequencies (VanRaden 2008)
- or something else

By an argument similar to the previous one, estimates of $\widehat{\boldsymbol{a}}$ are identical across all possible « centerings », but EBVs are shifted by a constant which is a function of $(\mathbf{p}^{*(\mathbf{1})} - \mathbf{p}^{*(\mathbf{2})})'\widehat{\boldsymbol{a}}$.

Coding: « centering »

Genotype	101 Coding	012 Coding	Centered coding
aa	$-a_i$	0	$-2p_i a_i$
Aa	0	a_i	$(1-2p_i)a_i$
AA	a_i	$2a_i$	$(2-2p_i)a_i$

To obtain correct results, *again*, one should be coherent and use the same coding $\boldsymbol{Z}^* = \mathbf{Z_{012}} - \mathbf{2p}^{*\prime}$ in all steps

For instance if you do SNP-BLUP with 10,000 animals and e.g. use observed allele frequencies (« centered coding ») then you MUST use the same frequencies for 100 newly genotyped animals, and not computing frequencies again

Unfortunately many packages (including blupf90) « center » by default without the user knowing exactly what happens. *Do Read the output on screen and the manual.*

Coding: « centering »

How do animal breeders work?

- run a <u>SNP-BLUP</u> periodically (say 3/year),
  - store reference alleles,
  - define and store $\mathbf{p}^*$,
  - build $\boldsymbol{Z}^* = \mathbf{Z_{012}} - \mathbf{2p}^*$
  - obtain $\widehat{\boldsymbol{a}}$
  - compute $\boldsymbol{Z}^*\widehat{\boldsymbol{a}}$ -> GEBVs

- In between SNP-BLUPs: do <u>Indirect Predictions</u> as
  - read reference alleles,
  - read $\mathbf{p}^*$,
  - read $\widehat{\boldsymbol{a}}$
  - build $\boldsymbol{Z}^* = \mathbf{Z_{012}} - \mathbf{2p}^*$
  - compute $\boldsymbol{Z}^*\widehat{\boldsymbol{a}}$ -> GBEVs

Coding: « centering »

How do animal breeders using blupf90 suite work (using defaults)?
- run a (ss)G<u>BLUP</u> periodically (say 3/year),
  - *blupf90*:
    - reference alleles are handled by the user (file is read as 012),
    - compute observed $\mathbf{p}^*$,
    - build $\boldsymbol{Z}^* = \mathbf{Z_{012}} - \mathbf{2p}^*$
    - use the equivalent model ssGBLUP and obtain GEBVs $\widehat{\boldsymbol{u}}$
  - *postGSf90*:
    - backsolve for marker effects $\widehat{\boldsymbol{a}} = \boldsymbol{f}(\boldsymbol{Z}^*, \widehat{\boldsymbol{u}})$
    - store $\widehat{\boldsymbol{a}}$
    - store $\mathbf{p}^*$
- In between SNP-BLUPs: do <u>Indirect Predictions</u>
  - *predf90*
    - reference alleles are handled by the user (file is read as 012),
    - read $\mathbf{p}^*$,
    - read $\widehat{\boldsymbol{a}}$
    - build $\boldsymbol{Z}^* = \mathbf{Z_{012}} - \mathbf{2p}^*$
    - compute $\boldsymbol{Z}^*\widehat{\boldsymbol{a}}$ -> GBEVs

Theoretical individual Reliabilities from SNP-BLUP

Reliability= $r^2(u_i, \hat{u}_i) = 1 - \frac{Var(\hat{u}_i)}{Var(u_i)} = 1 - \frac{Var(u_i|\boldsymbol{y})}{Var(u_i)}$

$\hat{u}_i = \boldsymbol{z}_i \hat{\boldsymbol{a}}_i$; see details in the notes

- $Rel_i = 1 - \frac{Var(\hat{u}_i)}{Var(u_i)} = 1 - \frac{Var(\hat{u}_i)}{\boldsymbol{z}_i \boldsymbol{z}_i' \sigma_{a0}^2} = 1 - \frac{\boldsymbol{z}_i \boldsymbol{C}^{aa} \boldsymbol{z}_i'}{\boldsymbol{z}_i \boldsymbol{z}_i' \sigma_a^2}$
- $\boldsymbol{C}^{aa}$= chunk of the SNP part of the MME^{-1} describing the Prediction Error Variance of marker estimates
- This says that an individual is accurately predicted if its $\boldsymbol{z}_i$ carries more weight (1-2 rather than 0) in the markers that are better predicted
- which shows that animals need to be well connected to the reference population

Individual reliabilities from SNP-BLUP

- $Rel_i = 1 - \frac{Var(\hat{u}_i)}{\boldsymbol{z}_i \boldsymbol{z}_i' \sigma_a^2} = \frac{\boldsymbol{z}_i \boldsymbol{C}^{aa} \boldsymbol{z}_i'}{\boldsymbol{z}_i \boldsymbol{z}_i' \sigma_a^2}$
  - $Var(\hat{u}_i)$ can be obtained by sampling (Gibbs) or inversion
  - $\hat{u}_i$ and $Var(\hat{u}_i)$ are invariant to coding but…
  - $\boldsymbol{z}_i \boldsymbol{z}_i' \sigma_a^2$ is <u>not</u> invariant to coding
- Reliabilities depend on coding !!
  - Solution: define a contrast from some “base” population (Tier et al., 2018 WCGALP; Bermann et al., 2022 WCGALP)

Coding: « scaling »

Another method « centers and scales » , i.e. for each marker

$$\mathbf{z}^* = \frac{(\mathbf{z_{012}} - mean(\mathbf{z_{012}}))}{sd(\mathbf{z_{012}})} = \frac{(\mathbf{z_{012}} - 2p^*)}{\sqrt{2p^*(1-p^*)}}$$

because

p^*= observed frequency = half mean of $\mathbf{z}^*$ = $\mathbf{1}'\mathbf{z}^*/2n$

$sd(\mathbf{z_{012}})$= $2p^*(1-p^*)$

doing this is complicated because

- for very small p^* we obtain very large $\mathbf{z}^*$
- the p^* and $\mathbf{z}^*$ changes from run to run and we have shifts of mean
- heritabilities implicitly change! (we may see this later)

- I generally DO NOT recommend using « centered and scaled »

SNP-BLUP parameters

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{I}\sigma_{a0}^{-2} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}$$

let's check this now

SNP-BLUP parameters

How do we get the <u>variance of SNP effects</u>, σ_{a0}^2 ?

- You can estimate it (Bayes C, REML)
- Few software available (GenSel, GS3, probably BGLR)
- (again) Strandén and Christensen (2011) proved that the estimate of σ_{a0}^2 in a « SNP-REML » or « Bayes C » is invariant to « choice of reference alleles » of **Z** and to « shifting » **Z**

SNP-BLUP parameters

How do we get the <u>variance of SNP effects</u>, σ_{a0}^2 ?

- You can « guess » from the <u>genetic variance</u> σ_u^2
- Assume that you estimated (with pedigree and records, by REML) a genetic variance σ_u^2 . This variance refers to the pedigree base population (usually old one)

- How much genetic variation does each marker contribute? Assuming Hardy-Weinberg
  - SNP 1 contributes $2p_1q_1a_1^2$ to the genetic variance
  - SNP 2 contributes $2p_2q_2a_2^2$ to the genetic variance
  - …
  - $\sigma_u^2 = 2\sum p_iq_ia_i^2 \approx 2(\sum p_iq_i) \times \overline{(a_i^2)} \approx 2(\sum p_iq_i)\sigma_{a0}^2$
  - the last step assumes independent a_i and p_i and uses $Var(xy) = Var(x)Var(y)$, Bohrnstedt, G. W., & Goldberger, A. S. (1969). JASA, 64(328), 1439-1442
  - the assumptions works quite well

SNP-BLUP parameters

- Reversing the expression $\sigma_u^2 \approx 2(\sum p_i q_i)\sigma_{a0}^2$ gives
  - $\sigma_{a0}^2 \approx \frac{\sigma_u^2}{2(\sum p_i q_i)}$

- So, from « old » estimates of genetic variance and allele frequencies we have a figure for σ_{a0}^2
- Because σ_u^2 is the variance in the base population, then p_i should ideally be the allelic frequency base population – which are usually NOT genotyped. This is a continuous source of misunderstanding.
- Experience shows that the error made using *observed* (current) p_i instead of base population p_i is not too high

SNP-BLUP parameters

It is tempting to use estimated SNP effects $\hat{a}_i$ to estimate the genetic variance as $2\sum p_i q_i \hat{a}_i^2$, but it doesn't work:

- $\sigma_u^2 \ll 2\sum p_i q_i \hat{a}_i^2$

Estimated SNP effects are shrunken towards the mean and the figure $2\sum p_i q_i \hat{a}_i^2$ is much smaller than σ_u^2

If this worked, we wouldn't need REML ☺. We'd just run BLUP and compute crossproducts of EBVs

Not all p's are equal !!

Note that we have used p in two places

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\,\mathbf{Z} + \mathbf{I}\sigma_{a0}^{-2} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}$$

define and store $\mathbf{p}^*$, build $\boldsymbol{Z}^* = \mathbf{Z}_{012} - 2\mathbf{p}^*$

$$\sigma_{a0}^2 \approx \frac{\sigma_u^2}{2(\sum p_i q_i)}$$

Here we can use anything in $\mathbf{p}^*$!!

Here we have to use p_i as close as possible to allele frequencies in the base population for which σ_u^2 was estimated

The two $\mathbf{p}^*$ and p_i don't need to match !!

- All this is not much used
- But it prepares the terrain for GBLUP and ssGBLUP

GBLUP and G matrices

- GBLUP from SNP-BLUP
- GBLUP and genomic relationships

GBLUP from SNP-BLUP

- We have defined breeding values as sum of SNP effects: $\boldsymbol{u} = \boldsymbol{Za}$
- To refer breeding values to an average value of 0, we center using $-2p_i$. We can use:
  - allele frequencies p_i in the <u>pedigree base</u> population: then breeding values refer to the pedigree base population and we use the same scale as "regular" BLUP with **A**
  - (observed) allele frequencies p_i in the <u>genotyped</u> population: then breeding values refer to the genotyped population and we use a different scale as "regular" BLUP with **A** (BV are shifted)
  - this is another source of confusion ☹

Genotype	101 Coding	012 Coding	Centered coding
aa	$-a_i$	0	$-2p_i a_i$
Aa	0	a_i	$(1-2p_i)a_i$
AA	a_i	$2a_i$	$(2-2p_i)a_i$

GBLUP from SNP-BLUP

- We have defined breeding values as sum of SNP effects:
$$\boldsymbol{u} = \boldsymbol{Za}$$
- Because $Var(\boldsymbol{a}) = \boldsymbol{I}\sigma_a^2$, then
$$Var(\boldsymbol{u}) = \boldsymbol{Z}(\boldsymbol{I}\sigma_a^2)\boldsymbol{Z}' = \boldsymbol{ZZ}'\sigma_{a0}^2$$
- But before, we found out that $\sigma_a^2 = \frac{\sigma_u^2}{2\sum p_i q_i}$
  - where σ_u^2 and p_i refer to the same population (usually the pedigree base population).
- Substituting:
$$Var(\boldsymbol{u}) = \frac{\boldsymbol{ZZ}'}{2\sum p_i q_i}\sigma_u^2$$
- Finally, we factorize σ_u^2

VanRaden's "first G"

$$\boldsymbol{G} = \frac{\left(\boldsymbol{M}-2\boldsymbol{p}^{*\prime}\right)\left(\boldsymbol{M}-2\boldsymbol{p}^{*\prime}\right)'}{2\sum p_i q_i} = \frac{\boldsymbol{Z}\boldsymbol{Z}'}{2\sum p_i q_i}$$

VanRaden's "first G"

Genotypes {0,1,2}

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

$$\boldsymbol{G} = \frac{\left(\boldsymbol{M}-2\boldsymbol{p}^{*\prime}\right)\left(\boldsymbol{M}-2\boldsymbol{p}^{*\prime}\right)'}{2\sum p_i q_i} = \frac{\boldsymbol{Z}\boldsymbol{Z}'}{2\sum p_i q_i}$$

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

They don't need to be the same allele frequencies !! (but they usually are)

VanRaden's "first G"

- If I want my EBVs to be in the same scale as pedigree analyses the numerator should contain « base population allele frequencies »
- If I use observed allele frequencies, then the EBVs will be shifted by negative constant (= to the genetic progress)
- The difference can be large for selected traits

Genotypes {0,1,2}

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

$$\boldsymbol{G} = \frac{\left(\boldsymbol{M}-2\boldsymbol{p}^{*\prime}\right)\left(\boldsymbol{M}-2\boldsymbol{p}^{*\prime}\right)'}{2\sum p_i q_i} = \frac{\boldsymbol{Z}\boldsymbol{Z}'}{2\sum p_i q_i}$$

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

- If I want to use « old » σ_u^2 from pedigree analyses then the denominator should be « base population allele frequencies »
- If I use observed allele frequencies, then the denominator is « too small »
- In practice the difference is small

GBLUP

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

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

- when we started showing this *circa* 2009 people made analogies with "A-BLUP"
- is there anyone still using "A-BLUP"?

GBLUP

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{W} \\ \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{W} + \boldsymbol{G}^{-1}\sigma_u^{-2} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{pmatrix}$$

- 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

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{W} \\ \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{W} + \boldsymbol{G}^{-1}\otimes\boldsymbol{G}_0^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{pmatrix}$$

$\boldsymbol{G}_0$ is the matrix of genetic covariance across traits

usually $\boldsymbol{R} = \boldsymbol{I}\otimes\boldsymbol{R}_0$, where $\boldsymbol{R}_0$ 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

Again,

- Rel_i is NOT invariant to the allele frequencies used in $\mathbf{Z} = \boldsymbol{M} - \mathbf{2}\boldsymbol{p}^{*\prime}$
- A solution is to define a contrast

GREML, G-Gibbs…

Use of **G** to estimate variance components (heritability)…
It can be done with `blupf90+, gibbsf90+, AsReml, TM`…

The result will refer to an ideal population with whatever allelic frequencies we introduced in the *denominator* of $\boldsymbol{G} = \frac{\boldsymbol{Z}'\boldsymbol{Z}}{2\sum p_i q_i}$.

If you put observed allele frequencies then you refer the estimate of variance components to the « observed » population
If you put base allele frequencies you refer to the « old » population

In livestock *with large and good data bases* the difference is small
For a method to compare genetic variances across different G's, A, etc etc relationships, check Legarra, TPB 2016

GBLUP == SNPBLUP

- Both give the same solutions
  - (up to the small detail of "tuning" and"blending" so that actually $\boldsymbol{G}^{**} \Leftarrow (1-\propto)(a+b\mathbf{G})+\propto \boldsymbol{A}_{22}$; this is taken care of in blupf90)
- We can jump from SNP-BLUP to GBLUP

$$\hat{\boldsymbol{u}} = \boldsymbol{Z}\hat{\boldsymbol{a}}$$

- We can jump from GBLUP to SNP-BLUP

$$\hat{\boldsymbol{a}} = \frac{1}{2\sum p_i q_i}\boldsymbol{Z}'\boldsymbol{G}^{-1}\hat{\boldsymbol{u}}$$

More gory stuff

- « Blending » -> making **G** invertible & accounting for genetic variance unexplained by markers
- « Tuning » -> making **G** similar to **A**

Tuning

- Having "base population allele frequencies" to get (on one hand) $2\sum p_i q_i$ and (on the other hand) $\mathbf{Z} = \boldsymbol{M} - \mathbf{2}\boldsymbol{p}^{*\prime}$ is "tout bénef" (all good)
  - Your genetic variance is on the right scale
  - Your EBVs are on the right scale
  - In other words, $\boldsymbol{G}$ and $\boldsymbol{A}_{22}$ are "comparable"
- USDA/CDCB have, for dairy, DNA samples from 1970's and can get base allele frequencies…
- most people don't
- …dozens of papers on "compatibility"

Tuning

- When base allele frequencies are <u>not</u> available there are 3 manners of "making compatible" $\boldsymbol{G}$ and $\boldsymbol{A}_{22}$
- Fix statistics of $\boldsymbol{G}$ so that they resemble those of $\boldsymbol{A}_{22}$ -> "tuning" **G**: Vitezica et al., 2011; Christensen et al., 2012
  - fixes both mean and variances
  - difficult to extend to several base population
- Add an intercept to account for the difference -> Fernando et al., 2014 "J factors"
  - fixes only means
  - can be extended to several base populations
  - only works in ssGBLUP
- Define a "new" base population with $\boldsymbol{p} = \mathbf{0.5}$ and "complete" pedigree relationships in $\boldsymbol{A}$: "metafounders" (Christensen 2012; Legarra et al., 2014)
  - fixes both means and variances
  - can be extended to several base populations

Tuning: Mean

- Fix statistics of $\boldsymbol{G}$ so that they resemble those of $\boldsymbol{A}_{22}$
- First, referring BV "from $\boldsymbol{G}$" to the same base as BVs "from $\boldsymbol{A}_{22}$"
  - In fact we introduce a *random* mean which compensates for $\boldsymbol{p}^* - \boldsymbol{p}_{base}$
  - because the mean is random, we can just add it, as a constant, to $\boldsymbol{G}$
  - $\boldsymbol{Var}(\mathbf{1}\mu + \boldsymbol{u}) = \mathbf{1}\mathbf{1}'Var(\mu) + \boldsymbol{G} = a + \boldsymbol{G} \Rightarrow \boldsymbol{G}^*$
  - It can be worked out that $a = Var(\mu) = mean(\boldsymbol{A}_{22}) - mean(\mathbf{G})$
  - If $\boldsymbol{G}$ is constructed with observed allele frequencies $a = mean(\boldsymbol{A}_{22}) \approx 2\bar{F}_p$ for $\bar{F}_p$ average pedigree inbreeding

Tuning: Variance

- Fix statistics of $\boldsymbol{G}$ so that they resemble those of $\boldsymbol{A}_{22}$
- Second, referring BV "from $\boldsymbol{G}$" to the same genetic variance as BVs "from $\boldsymbol{A}_{22}$"
  - In fact we introduce a scaling factor b which compensates for $2\sum p_i q_i - 2\sum p_i^* q_i^*$ which is the loss of heterozygosity from the base population to the one with p_i^*
  - using expectation theory $b = \frac{1+\overline{F_p}-\overline{\mathbf{A}}_{22}}{1+\overline{F_g}-\overline{\mathbf{G}}}$
  - $b\boldsymbol{G} \Rightarrow \boldsymbol{G}^*$
  - If $\boldsymbol{G}$ is constructed with observed allele frequencies $b = 1 + \bar{F}_p - \overline{\mathbf{A}}_{22} \approx 1 - \bar{F}_p$ for $\bar{F}_p$ average pedigree inbreeding
  - This corresponds exactly with the theory: the reduction in genetic variation is $1 - \bar{F}_p$

Tuning: Mean and Variance

- $\boldsymbol{G}^* \Leftarrow a + b\mathbf{G}$, a and b from previous slides
- Equivalently, you can get both numbers using two equations

$$mean\big(diag(\boldsymbol{G})\big)b + a = mean\big(diag(\boldsymbol{A}_{22})\big)$$
$$a + b\overline{\boldsymbol{G}} = \overline{\boldsymbol{A}_{22}}$$

- This is the strategy of Christensen et al. 2012
- Concepts are the same, and in practice it results in the same results as before
- This is the default in blupf90

Tuning: Mean and Variance

- Note
- All this works because adding a positive constant to a matrix keeps its "positive-definiteness"
- In practice, it means that if **G** is to be tuned, one needs to use in

$$\boldsymbol{G} = \frac{\left(\boldsymbol{M} - 2\boldsymbol{p}^{*'}\right)\left(\boldsymbol{M} - 2\boldsymbol{p}^{*'}\right)'}{2\sum p_i q_i} = \frac{\boldsymbol{Z}\boldsymbol{Z}'}{2\sum p_i q_i}$$

- allele frequencies that result in a and b being positive, otherwise the final $\boldsymbol{G}$ *may* not be invertible
- The right choice is observed frequencies or « close to base » (estimates of) allele frequencies

Blending

- **G** is often no invertible (clones, « centering », more individuals than markers)
  - However **G** is semi-positive definite.
  - We want invertible **G** to use in the MME
  - A practical solution is to « blend » **G** with a positive definite matrix to yield a modified invertible $\boldsymbol{G}^*$
- Blend with identity: $\boldsymbol{G}^* = (1-\alpha)\boldsymbol{G} + \alpha\boldsymbol{I}$ for α a small number, e.g. 0.01
- Blend with pedigree relationships: $\boldsymbol{G}^* = (1-\alpha)\boldsymbol{G} + \alpha\boldsymbol{A}_{22}$

Blending

- This has an extra interpretation
- Blend with pedigree relationships: $\boldsymbol{G}^* = (1-\alpha)\boldsymbol{G} + \alpha\boldsymbol{A}_{22}$

Fraction of genetic variation explained by markers

Fraction of genetic variation explained by pedigree

- In theory you can estimate α by REML
- In practice people use defaults (0.05 in blupf90) or do some cross-validation to find « the best α » (I think this is a bad idea)

Blending & Tuning: yet one slide

- We should mix only things that can be properly compared
- The right manner to work is
  1. Tuning: I adjust $\boldsymbol{G} = \frac{\boldsymbol{zz'}}{2\sum p_i q_i}$ to be similar to $\overline{\boldsymbol{A}_{22}}$: $\boldsymbol{G}^* \Leftarrow a + b\mathbf{G}$
  2. Blending: I « blend » with $\boldsymbol{A}_{22}$ to make $\boldsymbol{G}^*$ invertible: $\boldsymbol{G}^{**} \Leftarrow (1 - \propto)\boldsymbol{G}^* + \propto \boldsymbol{A}_{22}$
- Blupf90 did in the opposite order until ~2021 but this has been fixed now
  - There are no major consequences (but it's better to have everything right, you never know)

Single Metafounder

- Define a "new" base population with $\boldsymbol{p} = \mathbf{0.5}$ and "complete" pedigree relationships in $\boldsymbol{A}$: "metafounders" (Christensen 2012; Legarra et al., 2014)
  - fixes both means and variances
  - can be extended to several base populations

Single Metafounder

- Christensen (2012) suggests fitting **A** to **G** instead of the opposite
  - **A** depends on pedigree completion
  - Pedigrees are never complete !!
  - Ancestral relationships that can be seen in **G** go undetected in **A**
- Christensen analitically integrates out p_i (=allele frequencies) in a model that
  - uses $p = 0.5$ as reference in ALL loci and builds $\boldsymbol{G}_{05} = \frac{(\boldsymbol{M}-\mathbf{11}')(\boldsymbol{M}-\mathbf{11}')'}{2\sum 0.5\times 0.5}$
  - uses a relationship matrix $\mathbf{A}^{\gamma}$ with <u>related</u> founders
  - The parameter γ is the relationship across founders such that we see "current" genomic relationships

Single Metafounder

Classically we assume for founders

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

- Christensen changes this into:

$$\mathbf{A}^{\gamma} = \begin{pmatrix} 1+\frac{\gamma}{2} & \gamma & \gamma & \gamma \\ \gamma & 1+\frac{\gamma}{2} & \gamma & \gamma \\ \gamma & \gamma & 1+\frac{\gamma}{2} & \gamma \\ \gamma & \gamma & \gamma & 1+\frac{\gamma}{2} \end{pmatrix}$$

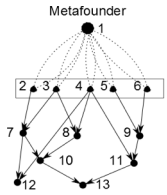
We proved that the same can be achieved definig an ancestor (a metafounder) that represents the base population and its average relationship (as referred to a population where all markers had $p = 0.5$) is γ

Metafounder relationships

RELATIONSHIPS

Across founders within the population

A SINGLE METAFOUNDER



Pedigree
1 0 0
2 1 1
3 1 1
4 1 1
5 1 1
6 1 1
7 2 3
8 3 4
9 5 6
10 7 8
11 4 6
12 7 4
13 10 11

$$\mathbf{A}^{\gamma} = \begin{pmatrix} \gamma & \gamma & \gamma & \gamma \\ \gamma & 1+\frac{\gamma}{2} & \gamma & \gamma \\ \gamma & \gamma & 1+\frac{\gamma}{2} & \gamma \\ \gamma & \gamma & \gamma & 1+\frac{\gamma}{2} \end{pmatrix}$$

It has self-relationship $A_{11} = \gamma$ so $F = \gamma - 1$.
If $\gamma = 0$ then we have regular relationships.
All **A** and $\mathbf{A}^{-1}$ methods work.

Single Metafounder

- Interestingly , *if we knew* base population frequencies p_i
$$\gamma = \sum(2p_i - 0.5)^2$$
  - which is the same as $\boldsymbol{G}_{05}$ with fractional genotypes
- For a single base population, the estimation of γ can be done by Maximum Likelihood comparing $\boldsymbol{G}_{05}$ and $\boldsymbol{A}_{22}$
- The "reference" genetic variance has changed – we need to scale genetic variances by $\sigma^2_{u(related)} \Leftarrow \frac{\sigma^2_{u(related)}}{\left(1-\frac{\gamma}{2}\right)}$
- The method can be extended to more populations – see later.

But what are genomic (additive) relationships?

Interpretation of **G**

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'**.

parenté
n. f.

1 Liens qui unissent les membres d'une famille. *Quel est votre lien de parenté avec elle? – C'est ma sœur.*

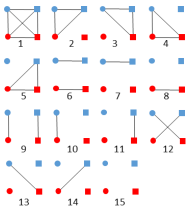
It obviously comes from Latin "parentes"

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

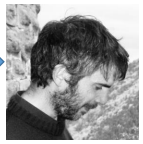


Relationships

- Relationships were conceived as standardized covariances (Fisher, Wright)
  - $Cov(u_i, u_j) = R_{ij}\sigma_u^2$
  - R_{ij} "some" relationship
  - σ_u^2 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



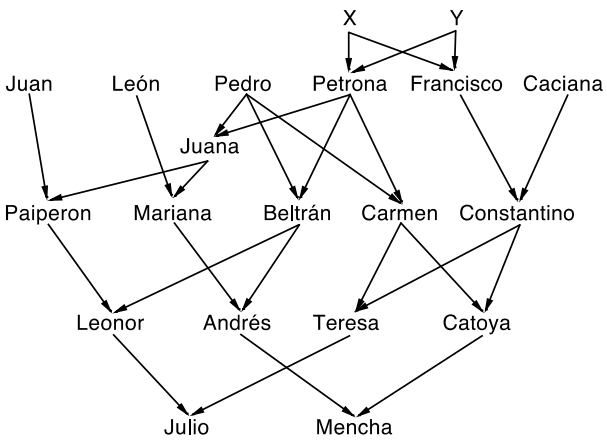


Figure 2. The pedigree of the Jicaque Indians Julio and Mencha.

	Julio-Mencha	Progeny-progeny	Julio-progeny	Mencha-progeny
δ_1	0.01025	0.06580	0.03467	0.03467
δ_2	0.02393	0.04333	0.08252	0.00000
δ_3	0.02490	0.04333	0.00000	0.08252
δ_4	0.02393	0.04333	0.06665	0.06665
δ_5	0.02490	0.04333	0.08228	0.08228
δ_6	0.00708	0.00729	0.00000	0.00000
δ_7	0.05103	0.02383	0.00000	0.00000
δ_8	0.05103	0.02383	0.00000	0.00000
δ_9	0.05127	0.24713	0.08228	0.06665
δ_{10}	0.10937	0.15900	0.29248	0.00000
δ_{11}	0.16992	0.15900	0.00000	0.29248
δ_{12}	0.00708	0.00729	0.06665	0.08228
δ_{13}	0.05103	0.02383	0.00000	0.29248
δ_{14}	0.05103	0.02383	0.29248	0.00000
δ_{15}	0.34326	0.08582	0.00000	0.00000

Garcia-Cortes Gen Sel Evol 2015

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 $\boldsymbol{A}$.
- $\boldsymbol{A}^{-1}$ 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

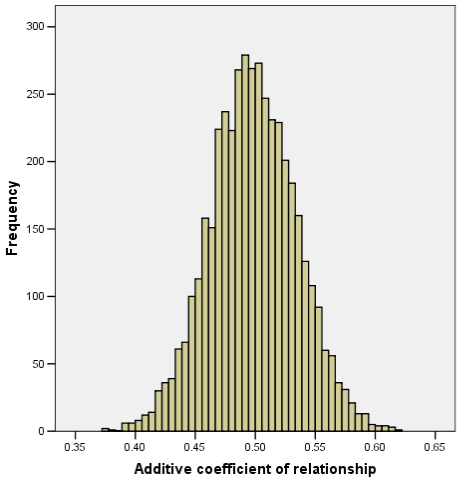
Early use of markers used them to infer A

- In conservation genetics, molecular markers have often been used to estimate pedigree relationships
- Gather markers, then reconstruct pedigrees, then construct **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 <u>expectation</u> of realized relationhips R
- **A** is <u>false</u>, and is « very » false for small values of **A**
- 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 »

Comparison of expected and observed variances – relationship/sharing



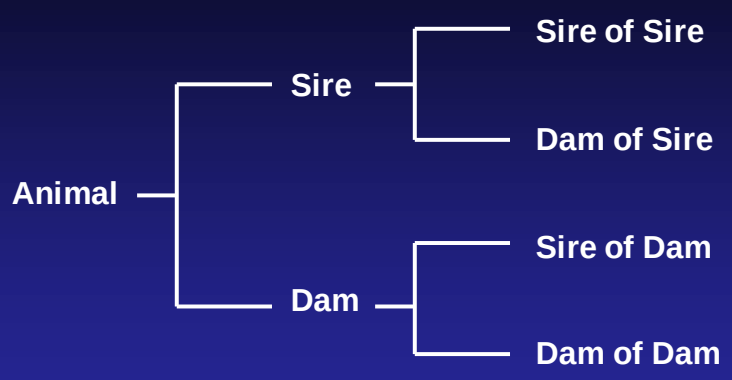
4401 full sib pairs
400-800 markers
Expected
Mean 0.5
SD 0.039

Observed
Mean 0.498
SD 0.036
Range 0.37 - 0.63

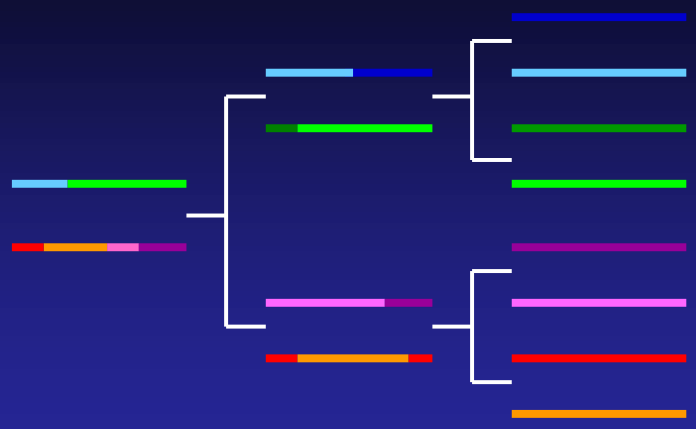
Source: Visscher et al.

Slide from WG Hill

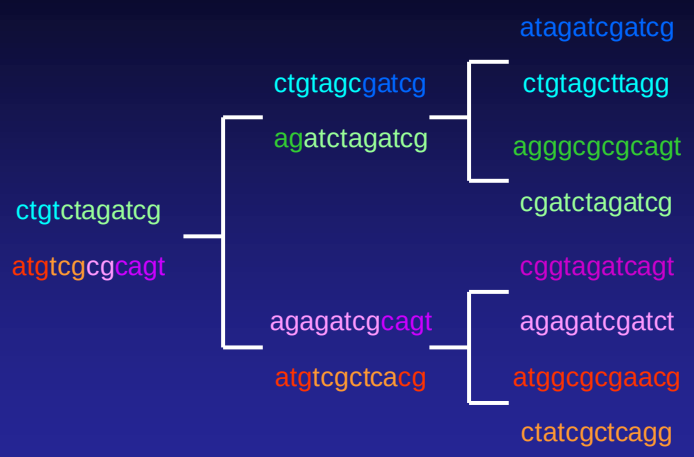
Traditional Pedigree



Genomic Pedigree

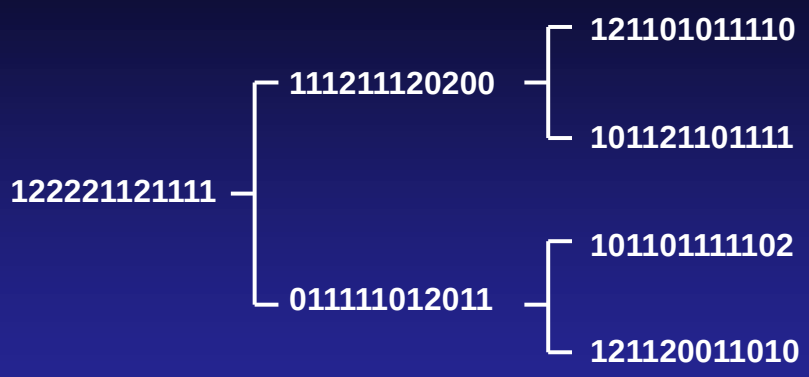


Haplotype Pedigree



Genotype Pedigree

Count number of second allele



0 = homozygous for first allele (alphabetically)
1 = heterozygous
2 = homozygous for second allele (alphabetically)

Covariance of gene content (seen as a trait)

- 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 $\hat{R}_{ij} = \frac{z_i z_j}{2pq}$
- Extended to many loci $\hat{R}_{ij} = \frac{mean(\mathbf{z}_i \mathbf{z}_j')}{mean(2\sum p_k q_k)} = \frac{\mathbf{z}_i \mathbf{z}_j'}{2\sum p_k q_k}$

VanRaden's "first G"

Genotypes {0,1,2}

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

If <u>base</u> allelic frequencies are used, **G** is an unbiased an efficient estimator of IBD realized relationships

If <u>observed</u> allelic frequencies are used, **G** is a biased (but accurate !!) estimator of IBD realized relationships

Some properties of G

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

$$\boldsymbol{G} = \frac{(\boldsymbol{M}-2\boldsymbol{P})(\boldsymbol{M}-2\boldsymbol{P})'}{2\sum p_i q_i}$$

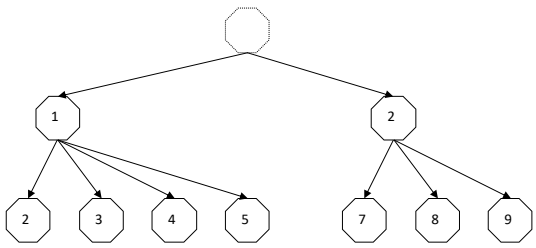
	AA	Aa	aa
freq	$q^2 + pqF$	2pq(1-F)	$p^2 + pqF$

Some intriguing properties of G

- If *p* are computed from the data
  - This implies that E(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 <u>wrong praxis</u>

Real results (AMASGEN)

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



Pedigree-based relationship

Little inbreeding

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

Cousin relationships ~0.125

"first G" genomic relationship

Less than 1 in the diagonal

Negative coefficients

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]
[1,]	0.82	0.40	0.43	0.38	0.12	0.04	0.04	0.01	0.10
[2,]	0.40	0.91	0.18	0.24	0.02	0.05	-0.04	-0.04	0.04
[3,]	0.43	0.18	0.88	0.19	0.07	0.00	0.07	-0.02	0.05
[4,]	0.38	0.24	0.19	0.86	0.02	-0.01	-0.02	0.01	0.03
[5,]	0.12	0.02	0.07	0.02	0.73	0.34	0.30	0.31	0.35
[6,]	0.04	0.05	0.00	-0.01	0.34	0.85	0.15	0.14	0.18
[7,]	0.04	-0.04	0.07	-0.02	0.30	0.15	0.80	0.14	0.17
[8,]	0.01	-0.04	-0.02	0.01	0.31	0.14	0.14	0.80	0.17
[9,]	0.10	0.04	0.05	0.03	0.35	0.18	0.17	0.17	0.85

$$\mathbf{G} = \mathbf{ZZ}'/2\sum_{all\ SNPs} p_i(1-p_i)$$ Relationships among cousins are ~0

IBS relationships at the markers

- $\boldsymbol{G}_{IBS}$ is a genomic relationship matrix based on Identity By State at the markers
- The terms in $\mathbf{G}_{IBS}$ are usually described in terms of identities or countings:
- $\boldsymbol{G}_{IBS_{ij}} = \frac{1}{n}\sum_{m=1}^{n} 2\,\frac{\sum_{k=1}^{2}\sum_{l=1}^{2} I_{kl}}{4}$,
  - where I_{kl} measures the identity across all 4 combinations of alleles

IBS and IBD

- IBS at markers (G_{IBSij}) is a frequently used estimator of realized IBD (R_{ij})
- Individuals can be identical by IBD or by IBS at the founders:

$$G_{IBSij} = R_{ij} + \left(2 - R_{ij}\right)(p^2 + q^2)$$

- Thus, IBS is biased upwards with respect to IBD.
- This has originated a bunch of estimators, with a common problem: where to get *p* from.
- For a detailed account, see Toro et al (2011 Gen Sel Evol)

GBLUP == GBLUP with IBS

- $\boldsymbol{G_{IBS}} = \frac{1}{2}\boldsymbol{G_{0.5}} + \boldsymbol{11}'$ where $\boldsymbol{G_{0.5}}$ is built pretending that $p = 0.5$
- The implicit denominator in $\boldsymbol{G_{IBS}}$ is "too big"
- Note that e.g. $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} 0.5 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} 1$,
- in other words, what matter is the product $\boldsymbol{G_{0.5}}\ \sigma^2_{u"0.5"}$
- we can scale the genetic variance appropriately as (n = number of markers)

$$\sigma^2_{u"0.5"} \Leftarrow \sigma^2_u \frac{n}{2\sum p_i q_i}$$

- Then we get the same GEBVs as with « normal **G** »

OK, so what allele frequencies should I use?

- If you "know" and want to use the variance components:
  - Try to use base allele frequencies
  - If not, use a "tuned" **G** or metafounders theory
- If you "don't know" variance components and you estimate them
  - If using REML or Bayesian methods you get the variance components just right
  - However, inferring variance components and h^2 gets tricky,
  - because they refer to the population with the allele frequencies in the denominator of **G**
  - See Legarra 2016, Theor Pop Biol

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

GWAS

- brief history of QTL detection
- GWAS from single marker regression or GWAS from GBLUP
- what GWAS signals mean

Brief history

- Geneticists always want to find genes, but it is a very difficult task
- 1989: Lander & Botstein propose a systematic scan using linkage and microsatellites
- These methods were based on following putative different alleles using microsatellites -> within-family linkage
- explossion of « QTL » studies in human, plant and livestock

A QTL with major effect on milk yield and composition maps to bovine Chromosome 14

Wouter Coppieters, Juliette Riquet, Juan-José Arranz, Paulette Berzi, Nadine Cambisano, Bernard Grisart, Latifa Karim, Fabienne Marcq, Laurence Moreau, Carine Nezer, Patricia Simon, Pascal Vanmanshoven, Danny Wagenaar, Michel Georges

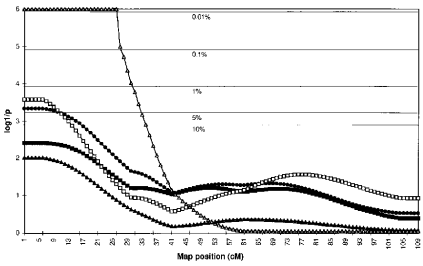
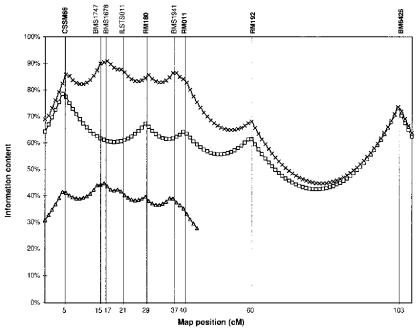


Fig. 2. Location scores obtained along the Chr 14 microsatellite map with the sum-of-rank QTL mapping method (Coppieters et al. 1998b). Evidence in favor of the presence of a QTL at a given map position is measured as *log(1/p)* with *p* corresponding to the p-value of the actual data as measured from chromosome-wise analyses of 10^6 phenotype permutations. Experiment-wise thresholds obtained by Bonferroni correction of the chromosome-wise thresholds to account for the analysis of three multiple traits and 29 chromosomes are indicated by horizontal bars. ●: milk yield; ■: protein yield; □: protein percentage; ▲: fat yield; △: fat percentage. For fat percentage (△) the location profiles are flat between positions 0 and 26cM because the χ^2 values obtained from the real data within this segment exceeded all values obtained from the 10^6 permutations.



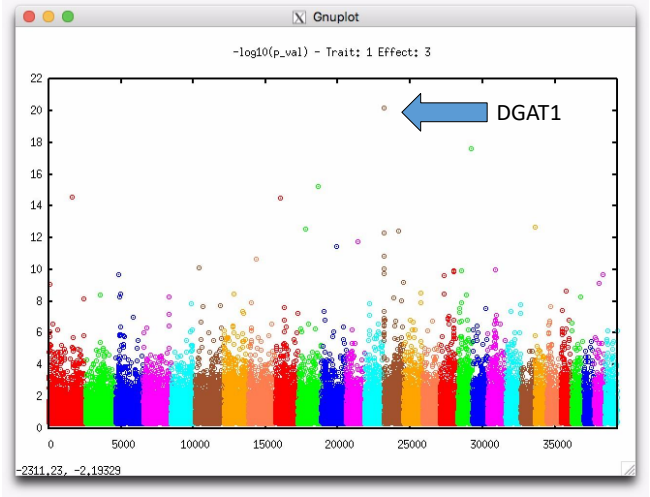
Microsatellite genotyping. A whole genome scan was undertaken with a battery of 215 microsatellite markers compiled from published marker maps and jointly covering 2947 cM (Haldane) of the bovine genome

Fig. 1. Information content of the used Chr 14 microsatellite map in the granddaughter design (□: low density map; x: high density map) and grand^2-daughter design (▲).

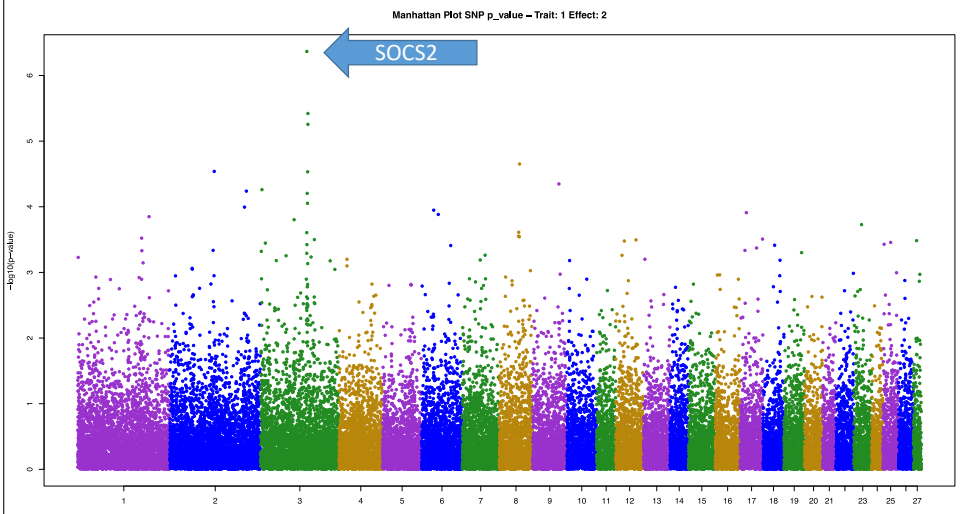
Methods: from (ss)GBLUP

- The use of high-density SNP chips shifted the methods towards marker-based regressions
- the hope is that the marker close(st) to the QTL is in linkage disequilibrium

Uruguayan Holstein



1000 Lacaune sheep



Methods: Single-marker regression

- Single-marker regression (« one marker at a time ») with accounting for relationships

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{z}\beta + \boldsymbol{Wu} + \boldsymbol{e}$$
$$Var(\boldsymbol{u}) = \boldsymbol{G}\sigma_u^2$$

- β effect of the marker; $\boldsymbol{z}$ incidence <u>vector</u> as we have seen
- test: $\frac{\widehat{\beta}}{s.e.(\widehat{\beta})} \sim N(0,1)$

Methods: from (ss)GBLUP

Meta-analysis of genome-wide association from genomic prediction models

Y. L. Bernal Rubio*†, J. L. Gualdrón Duarte*, R. O. Bates*, C. W. Ernst*, D. Nonneman‡, G. A. Rohrer‡, A. King‡, S. D. Shackelford‡, T. L. Wheeler‡, R. J. C. Cantet†§ and J. P. Steibel*¶

Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations

Jose L Gualdrón Duarte[1], Rodolfo JC Cantet[1], Ronald O Bates[2], Catherine W Ernst[2], Nancy E Raney[2] and Juan P Steibel[2,3*]

Genome-Wide Association Analyses Based on Broadly Different Specifications for Prior Distributions, Genomic Windows, and Estimation Methods

Chunyu Chen,[1] Juan P. Steibel, and Robert J. Tempelman
Department of Animal Science, Michigan State University, East Lansing, Michigan 48824
ORCID ID: 0000-0002-7833-6730 (R.J.T.)

GWAS by GBLUP: Single and Multimarker EMMAX and Bayes Factors, with an Example in Detection of a Major Gene for Horse Gait

Andres Legarra,*,1 Anne Ricard,†,‡ and Luis Varona§,**

Genome-wide association analyses based on a multiple-trait approach for modeling feed efficiency

Y. Lu,* M. J. Vandehaar,* D. M. Spurlock,† K. A. Weigel,‡ L. E. Armentano,‡ E. E. Connor,§ M. Coffey,# R. F. Veerkamp,‖ Y. de Haas,‖ C. R. Staples,¶ Z. Wang,** M. D. Hanigan,†† and R. J. Tempelman*1
*Department of Animal Science, Michigan State University, East Lansing 48824

Methods: from (ss)GBLUP

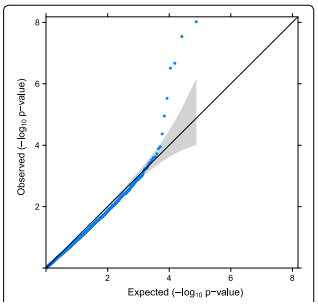
- It can be proved (see lengthy details in the Notes) that if $\boldsymbol{G} = \frac{\boldsymbol{Z}'\boldsymbol{Z}}{2\sum p_i q_i}$
- and we
  1. backsolve SNP effects using $\hat{\boldsymbol{a}}|\hat{\boldsymbol{u}} = \frac{1}{2\Sigma p_i q_i}\boldsymbol{Z}'\boldsymbol{G}^{-1}\,\hat{\boldsymbol{u}}$
  2. compute s.e. $(\hat{a})$
- Then $\frac{\hat{\beta}}{s.e.(\hat{\beta})} = \frac{\hat{a}}{s.e.(\hat{a})}$
- This makes our life easier: we just need to run a GBLUP and backsolve
  - implemented on blupf90+ , postGSf90
- The same can be obtained directly using SNP-BLUP

Details on GWAS

- the purpose of GWAS is (or should be) to inform on the *etiology* of diseases or causal mechanisms of traits
- It is better to find 2 very good hits than 50 small ones !!
- Choice of an indicator
  - I strongly recommend using p-values
    - If you want to detect a gene you need to be sure.
    - p-values are there to control how many mistakes (false positives) you will made
    - if you don't like thresholds, use FDR
  - effect sizes are VERY misleading:
    - more polymorphic markers have larger effects
    - small studies will have large effect just by chance

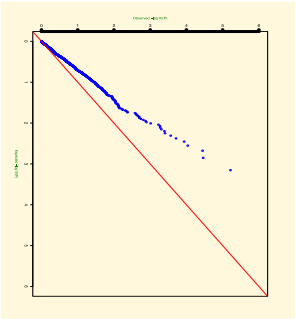
Details on GWAS: QQ plots

- If the model is correct, Qqplots should align properly. if not, they may not
- If you can't fix the model, then you can use "genomic control" (Devlin and Roeder, 1999, 2004) which is a hack that basically scales the p-values



Good

Bad



Non-additive effects in the genetic evaluation

Zulma G. Vitezica

INPT / INRAE - UMR GenPhySE, Toulouse, France

zulma.vitezica@toulouse-inp.fr

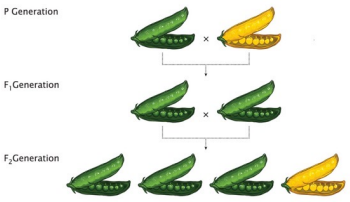
Plan

- Biological vs. statistical effects
  - Why statistical effects matter
- New models accounting for non-additive effects
  - GBLUP, GDBLUP, and its extensions
  - Inbreeding depression
- Is this any useful?
  - Extra accuracy in predictions
  - Variance components
  - Mate allocation
- Conclusions

Biological *vs.* statistical effects

Biological effects

The terms ‘dominance/epistasis’ describe apparent distortions of mendelian segregation ratios that were due to one gene masking the effects of another



Example of dominance

Genotypes at locus 1	Genotypes at locus 2		
	BB	Bb	bb
AA	🐴	🐴	🐴
Aa	🐴	🐎	🐴
aa	🐴	🐴	🐴

Example of epistasis: dominance-by-dominance two-locus epistasis

Biological effects

- Unfortunately **we don't know all gene actions & pathways**
- For many purposes, we need to make educated guesses
- Guesses include:
  - predicting phenotype of progeny (Genetic evaluation)
  - Is this genome region interesting? (GWAS)
  - What happened in this genome region? (selection footprints)
- For these practical purposes, we use statistical models

Statistical effects

- Fisher's described dominance and epistasis as deviations from additivity in a linear statistical model
- Statistical effects (dominance & epistasis) are a population phenomenon
- Genetic model

Phenotype

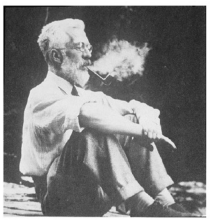
Epistatic deviation

$$P = A + D + I + E$$

Additive (or breeding) value

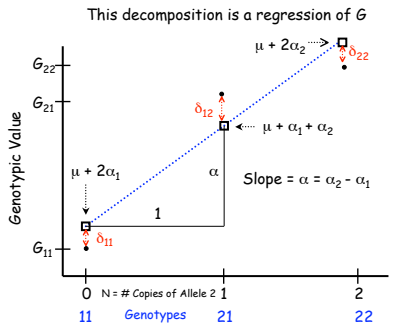
Dominance deviation

Statistical effects



Fisher (1918) explained that the substitution effect of one allele is the regression of phenotype on genotype

$$\alpha = (\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'\boldsymbol{y}, \quad \boldsymbol{z} = \begin{cases} \mathbf{0} \\ \mathbf{1} \\ \mathbf{2} \end{cases}$$

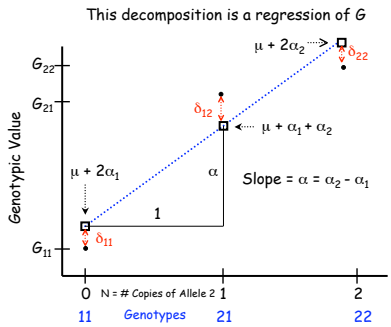


Walsh B., 2013

- Dominance deviations are essentially residuals
- Dominance deviations are the difference for a genotype (in red) between the genotypic value and its prediction from 2 alleles.

Statistical effects

- Why is α relevant & how does it take care of non-additive gene action?
  - The statistical definition doesn't care how α "works"
  - By definition, α potentially includes biological dominance and epistasis
  - Because individuals pass on gametes (and not complete genotypes) to offspring:
  - α describes how much you gain by selecting an allele (in either natural or artificial selection)



Walsh B., 2013

Example pairwise epistasis

$\alpha_1 =$	Biological
a_A	Additive
$+(1-2p_1)d_A$	Dominant
$+(p_2-q_2)i$	Additive x additive
$+(2p_2q_2)j$	Additive x dominant
$+(1-2p_1)(p_2-q_2)l$	Dominant x additive
$+2p_2q_2(1-2p_1)k$	Dominant x dominant

	BB	Bb	bb
AA	$y_1 = m + a_A + a_B + i$	$y_4 = m + a_A + d_B + j$	$y_7 = m + a_A - a_B - i$
Aa	$y_2 = m + d_A + a_B + l$	$y_5 = m + d_A + d_B + k$	$y_8 = m + d_A - a_B - l$
aa	$y_3 = m - a_A + a_B - i$	$y_6 = m - a_A + d_B - j$	$y_9 = m - a_A - a_B + i$

$$\mu = m + a_A(p_1 - q_1) + a_B(p_2 - q_2) + 2p_1q_1d_A + 2p_2q_2d_B + (p_1 - q_1)(p_2 - q_2)i + 2(p_1 - q_1)p_2q_2j + 2p_1q_1(p_2 - q_2)l + 4p_1q_1p_2q_2k$$

$$\alpha_1 = a_A + d_A(1\text{-}2p_1) + (p_2\text{-}q_2)i + 2p_2q_2j + (1\text{-}2p_1)(p_2\text{-}q_2)l + 2p_2q_2k(1\text{-}2p_1)$$

$$\alpha_2 = a_B + d_B(1\text{-}2p_2) + (p_1\text{-}q_1)i + 2p_1q_1l + (1\text{-}2p_2)(p_1\text{-}q_1)j + 2p_1q_1k(1\text{-}2p_2)$$

$$d_1 = (d_A - l) + 2p_2(l + k) - 2kp_2^2$$

$$d_2 = (d_B - j) + 2p_1(j + k) - 2kp_1^2$$

$$\alpha_1\alpha_2 = (i + j + k + l) - 2p_1(j + k) - 2p_2(j + k) - 2kp_1p_2$$

$$\alpha_1 d_2 = (j + k) - 2kp_1$$

$$d_1\alpha_2 = (1 + k) - 2kp_2$$

$$d_1 d_2 = k$$

Toro, 2017

Statistical & biological effects

Genotypes at locus 1	Genotypes at locus 2		
	BB	Bb	bb
AA			
Aa			
aa			

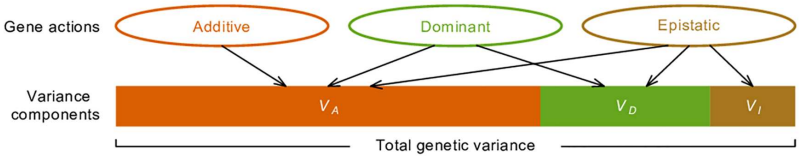
$\alpha_1 = 2p_2q_2(1 - 2p_1)($  -  $)$

Being a "Big" horse is determined by biological **dominance-by-dominance two-locus epistasis**

Genotypes at locus 1	Genotypes at locus 2		
	BB	Bb	bb
AA	μ	μ	μ
Aa	μ	$\mu + \{dd\}_{12}$	μ
aa	μ	μ	μ

Statistical & biological effects

- In the classical $V_A + V_D + V_I$ partition,
  - Additive biological gene actions contribute only to V_A, while
  - Both biological dominant and biological epistatic gene actions contribute to multiple variance components



Huang & Mackay, 2016

  - There is no correspondence between the kind of biological gene action and the variance component

What to do with all these math?

- In absence of knowing true action genes, this gives tools
- α (statistical additive effect) says
  - how much do you improve if you select me
  - Big α = interesting locus
- d_i^* (statistical dominance effect) says
  - For whatever reason, the heterozygote here is interesting
  - Perhaps we can mate these two animals here and maximize it
- $(\alpha\alpha)_{ij}$ (statistical epistatic effect) says
  - Somehow the fates of these two loci are bound together

What to do with all these math?

- α (statistical additive effect) is the ONLY component involved in selection, because only individual alleles are transmitted from parents to descendants
- d_i^* (statistical dominance effect) and $(\alpha\alpha)_{ij}$ (statistical epistatic effect) also contribute to the total genetic value and to the expected phenotype of the crosses/hybrid, but not to selection, because the allele/gene combinations are not transmitted to the descendants

New models accounting for non-additive effects

GBLUP, GDBLUP, and its extensions

Inbreeding depression

“Mixed model” based prediction

- We use quantitative genetic theory to build relationship matrices
- Then we fit them into mixed model

Genomic prediction with non-additive effects

1. We need to construct a linear model based on SNP genotypes
2. Write orthogonal incidence matrices for additive, dominant, additive x additive, additive x dominant… SNP effects
   1. This yields SNP-BLUP or RR-BLUP kind of models but they are cumbersome for epistasis
3. Equivalently, define relationship matrices
   1. High order matrices are products of low order matrices
   2. The whole theory stems from
      1. VanRaden 2008 (A),
      2. Vitezica et al., 2013 **(A+D)**
      3. Vitezica et al., 2017 (**A+D+AxA + any epistatic interactions**)
      4. González-Diéguez et al. (2021) (**A+D+AxA + any epistatic interactions in hybrid crops)**
4. Use a Mixed Model with relationship matrices

<u>This is doable if</u> all individuals are genotyped

- **There is no Single Step GBLUP for dominance or epistasis**

Genomic prediction with non-additive effects

- Recipe:
  1. Define incidence matrices **Z** for α and **W** for d^* , e.g.

$$Z_{ij} = \begin{cases} 2-2p \\ 1-2p \\ 0-2p \end{cases} \quad \text{and} \quad W_{ij} = \begin{cases} -2q^2 \\ 2pq \\ -2p^2 \end{cases} \text{ for genotypes } \begin{cases} AA \\ Aa \\ aa \end{cases}$$

  2. Relationship matrices are:
     - $\boldsymbol{G_A} = \frac{\boldsymbol{ZZ'}}{2\sum p_i q_i}$ for individual additive effects (GEBVs)
     - $\boldsymbol{D} = \boldsymbol{G}_D = \frac{\boldsymbol{WW'}}{4\sum (p_i q_i)^2}$ for dominance deviations

Use in Mixed Model: GD-BLUP

Genomic prediction with non-additive effects

- Recipe:
  2. Relationship matrices are:
     - $\boldsymbol{G_A} = \boldsymbol{ZZ'}/2\sum p_i q_i$ for individual additive effects (GEBVs)
     - $\boldsymbol{D} = \boldsymbol{G}_D = \boldsymbol{WW'}/4\sum (p_i q_i)^2$ for dominance deviations

     - $\boldsymbol{G_{AA}} = \boldsymbol{G_A} \odot \boldsymbol{G_A}/\mathrm{mean}(\mathrm{diag}(\boldsymbol{G_A} \odot \boldsymbol{G_A}))$ for additive x additive
     - $\boldsymbol{G_{AD}} = \boldsymbol{G_A} \odot \boldsymbol{G_D}/\mathrm{mean}(\mathrm{diag}(\boldsymbol{G_A} \odot \boldsymbol{G_D}))$ for additive x dominant
     - …
     - … e.g. $\boldsymbol{G_{AAD}} = \boldsymbol{G_A} \odot \boldsymbol{G_A} \odot \boldsymbol{G_D}/\mathrm{mean}(\mathrm{diag}(\boldsymbol{G_A} \odot \boldsymbol{G_A} \odot \boldsymbol{G_D}))$

Genomic prediction with non-additive effects

- Recipe:
  2. Relationship matrices are:
     - $\boldsymbol{G_{AD}} = \boldsymbol{G_A} \odot \boldsymbol{G_D} / \mathrm{mean}(\mathrm{diag}(\boldsymbol{G_A} \odot \boldsymbol{G_D}))$ for additive x dominant

Genomic **additive** relationship matrix

Genomic **dominant** relationship matrix

$$\boldsymbol{G}_{AD} = \frac{\boldsymbol{G}_A \odot \boldsymbol{G}_D}{tr(\boldsymbol{G}_A \odot \boldsymbol{G}_D)/n}$$

Use in Mixed Model: GDI-BLUP

A <u>standardization</u> based on the trace of the relationship matrices <u>is needed</u>.

Genomic prediction with non-additive effects

- Recipe:
  - Then use these matrices in (G)(D)(I)BLUP / REML

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{g}_A + \boldsymbol{g}_D + \boldsymbol{g}_{AA} + \boldsymbol{g}_{AD} + \boldsymbol{g}_{DD} + \cdots (+\boldsymbol{pe}) \ldots + \boldsymbol{e}$$

$$Var(\boldsymbol{g}_A) = \boldsymbol{G}_A \sigma_A^2; Var(\boldsymbol{g}_D) = \boldsymbol{D}\sigma_D^2; Var(\boldsymbol{g}_{AA}) = \boldsymbol{G}_{AA}\sigma_{AA}^2$$

- $\boldsymbol{pe}$ is the permanent environmental effect
  - captures remaining genetic effects (e.g. AxAxAxA…) in repeated records (such as analysis of milk yield)
- The matrices of higher orders $\boldsymbol{G}_{AA}, \boldsymbol{G}_{AAA}, \boldsymbol{G}_{AAAA}$ are increasingly less informative and at some point <u>they're not worth fitting</u>.

Genomic prediction with non-additive effects – crosses in hybrid crops

- In hybrid crops like maize, the cultivated plant is usually an F1 hybrid which is the cross of two homozygote lines, each from a different population ("heterotic group")
- Parental homozygote lines are homozygous at all loci
- This generates a particular partition of additive, dominance and epistasis *across* and *within* heterotic groups

Genomic prediction of hybrid crops allows disentangling dominance and epistasis

David González-Diéguez [1]*, Andrés Legarra[1], Alain Charcosset[2], Laurence Moreau[2], Christina Lehermeier [3], Simon Teyssèdre[3], and Zulma G. Vitezica [1]

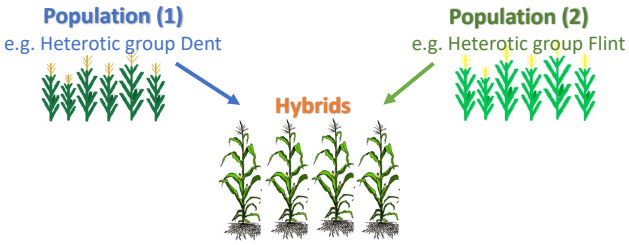
GENETICS, 2021, **218(1)**, iyab026

DOI: 10.1093/genetics/iyab026

Genomic prediction in hybrid crops

- Hybrid crops from pure lines
  - E.g. maize: population 1 is "Dent" and population 2 is "Flint"
  - The effects (GCA and SCA) are defined "according to parental origin"
  - We define **Z**-matrices <u>within each heterotic groups</u>
  - **W**-matrix is defined <u>in the hybrid</u>

David González-Diéguez



Genomic prediction in hybrid crops

- Recipe:
  1. For each locus,

  define incidence matrices $\mathbf{Z}_1$ for α_1 (pop 1), $\mathbf{Z}_2$ for α_2 (pop 2) and $\mathbf{W}$ for d^* (in hybrids)

$$Z_{1_{ij}} = \begin{cases} (1-p_1) \\ (\quad -p_1) \end{cases} \text{for genotypes} \begin{cases} B_1B_1 \\ b_1b_1 \end{cases}, \qquad Z_{2_{ij}} = \begin{cases} (1-p_2) \\ (\quad -p_2) \end{cases} \text{for genotypes} \begin{cases} B_2B_2 \\ b_2b_2 \end{cases}$$

and

$$W_{ij} = \begin{cases} -2q_1q_2 \\ 2q_1p_2 \\ 2p_1q_2 \\ -2p_1p_2 \end{cases} \text{for genotypes} \begin{cases} B_1B_2 \\ B_1b_2 \\ b_1B_2 \\ b_1b_2 \end{cases}$$

New models accounting for non-additive effects

GBLUP, GDBLUP, and its extensions

Inbreeding depression

Inbreeding / heterosis

- Inbreeding depression is the decline in biological fitness (viability, fertility, …) as a consequence of inbreeding

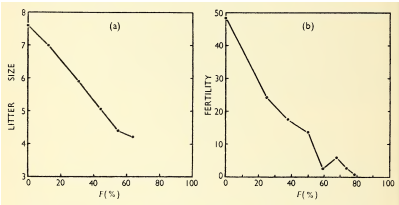


FIG. 14.1. Examples of inbreeding depression affecting fertility. (*a*) Litter-size in mice (original data). Mean number born alive in 1st litters, plotted against the coefficient of inbreeding of the litters.

(Falconer, 1981)

- This phenomenon may be explained by directional dominance.
- Directional dominance, e.g. the heterozygote is usually "better"

(Lynch & Walsh, 1998)

Inbreeding/ heterosis

- If heterosis or inbreeding depression, $E(\boldsymbol{d}) = \mathbf{1}\mu_D$ with $\mu_D > 0$
- Statistically this translates into a regression on a measure *F* of homozygosity ($\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{F}b + \boldsymbol{g}_A + \boldsymbol{g}_D + \cdots + \boldsymbol{e}$)
  - Across individual markers: "genomic inbreeding" (Silio et al 2013; Xiang et al 2016)
  - In blocks: ROHs (long ROHs are better because inbreeding has not been purged)
- Ignoring inbreeding/heterosis may inflate estimates of dominance variance
- Including inbreeding/heterosis allows finer estimates of EBV

Results?

OK, so we have this nice theory, what now?

- Is this any useful?
  - Extra accuracy in predictions
  - Variance components
  - Mate allocation

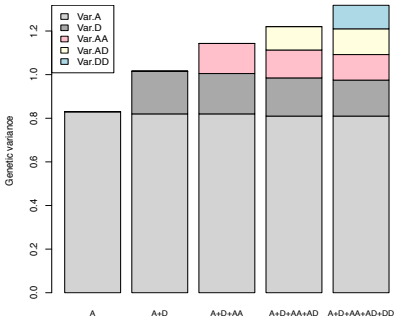
Example in pigs

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{F}b + \boldsymbol{g}_A + \boldsymbol{g}_D + \boldsymbol{g}_{AA} + \boldsymbol{g}_{DA} + \boldsymbol{g}_{DD} + \boldsymbol{pe} + \boldsymbol{e}$$

- Small variances for non-additive effects
- The model is empirically orthogonal: variance component estimates do not change by adding an extra term
- Inclusion of dominance/epistasis did not increase the accuracy of prediction of breeding values

Litter size
12.7 ± 3.1

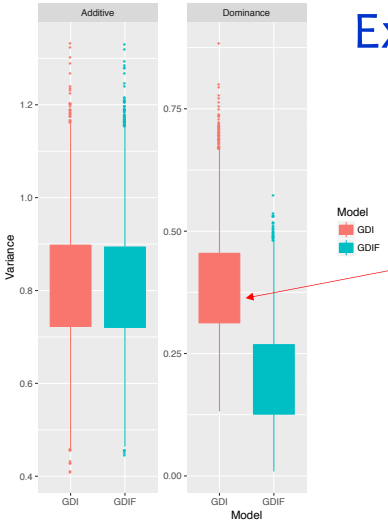
Genus plc (Hendersonville, TN, USA)
3,619 genotyped sows 13,369 records
38,779 SNPs



From Genus

Vitezica et al., 2018.

Example in pigs



Posterior distributions of additive and dominance genetic variances for model including (GDIF) or not (GDI) genomic inbreeding

From Genus

Without including inbreeding depression in the model, dominance variance was overestimated

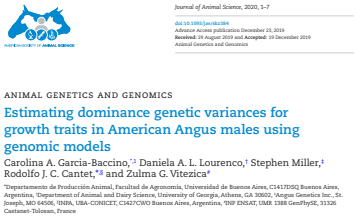
This has long been known for pedigree analysis (e.g. DeBoer and Hoeschele, 1993).

Vitezica et al., 2018.

Example in beef cattle

Carolina Garcia-Baccino

American Angus Association
19,375 genotyped males
39,245 SNPs



Journal of Animal Science, 2020, 1–7

doi:10.1093/jas/skz304
Advance Access publication December 23, 2019
Received: 20 August 2019 and **Accepted:** 19 December 2019
Animal Genetics and Genomics

ANIMAL GENETICS AND GENOMICS

Estimating dominance genetic variances for growth traits in American Angus males using genomic models

Carolina A. Garcia-Baccino,[1,2] Daniela A. L. Lourenco,[3] Stephen Miller,[4] Rodolfo J. C. Cantet,[1,6] and Zulma G. Vitezica[5]

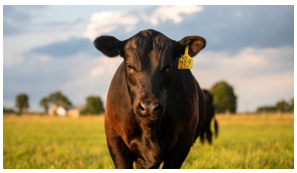
[1]Departamento de Producción Animal, Facultad de Agronomía, Universidad de Buenos Aires, C1417DSQ Buenos Aires, Argentina, [2]Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602, [3]Angus Genetics Inc., St. Joseph, MO 64506, [4]INRA, UBA-CONICET, C1427CWO Buenos Aires, Argentina, [5]INP ENSAT, UMR 1388 GenPhySE, 31326 Castanet-Tolosan, France

Small variances for non additive effects
Inclusion of dominance in the model did not increase the accuracy of prediction of breeding values

Table 2. Estimates of additive, dominance deviation, and residual variance components (σ^2_A, σ^2_D, σ^2_e) and heritability for growth traits using MG and MGD models

Trait[1]	Model[2]	σ^2_A	σ^2_D	h^2_A	h^2_D	(σ^2_D/σ^2_A)	σ^2_e
BW	MG	6.27 (0.33)	—	0.25	—	—	18.82 (0.24)
	MGD	6.28 (0.33)	0.18 (0.15)	0.25	0.01	0.03	18.65 (0.28)
WW	MG	222.75 (14.61)	—	0.16	—	—	1186.28 (14.26)
	MGD	223.55 (14.82)	10.02 (4.98)	0.16	0.01	0.04	1176.88 (14.86)
PWG	MG	270.76 (20.42)	—	0.16	—	—	1388.81 (19.87)
	MGD	270.30 (21.94)	21.68 (10.95)	0.16	0.01	0.08	1369.01 (26.00)

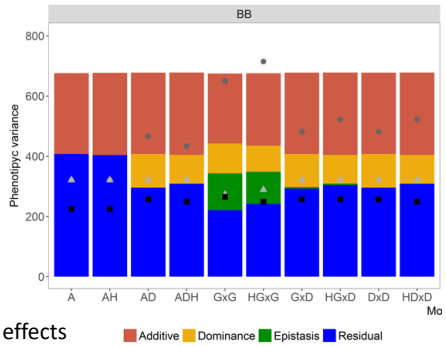
[1]BW, birth weight; WW, weaning weight; PWG, postweaning gain.
[2]MG, model including only additive effects; MGD, model including both additive and dominant effects.
The results are given as estimate (in parenthesis SE); $h^2_A = \sigma^2_A/\sigma^2_P$ and $h^2_D = \sigma^2_D/\sigma^2_P$, where σ^2_P is the phenotypic variance.

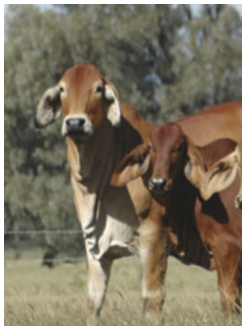


From AAA

Example in beef cattle

2,111 **Australian Brahman** (BB) cows and bulls
Genotyped with 770,000 SNPs
Body yearling weight





From ABBA

Small variances for non additive effects
Without including inbreeding depression in the model, dominance variance was overestimated

Raidan et al., 2018

Results

- Inclusion of dominance/epistasis
  - does not increase the accuracy of prediction of breeding values (Ertl et al., 2014; Xiang et al, 2016; Esfandyari et al., 2016; Moghaddar and van der Werf, 2017, González-Diéguez et al., 2019, Garcia-Baccino et al., 2020 – Pégard et al., 2020, González-Diéguez et al., 2021)
  - with the exception of Aliloo et al. (2016) (for fat yield in Holstein)
- Inclusion of inbreeding depression/heterosis effect
  - does increase predictive ability (Xiang et al., 2016) in pigs
  - and in maize (Roth et al., 2022)
- Fitting non-orthogonal models or non fitting inbreeding
  - Biases in variance component estimation (Vitezica et al. 2013; 2018)

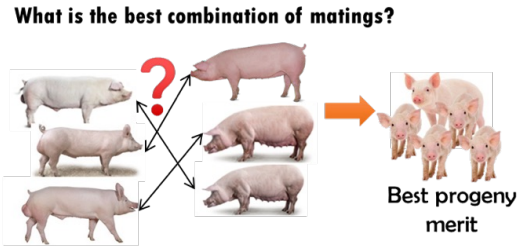
Results?

OK, so we have this nice theory, what now?

- Is this any useful?
  - Extra accuracy in predictions
  - Variance components
  - Mate allocation

Mate allocation: theory

- What happens if I mate *i* and *j* so that the product has an extraordinarily good phenotype (=dominance deviation)?

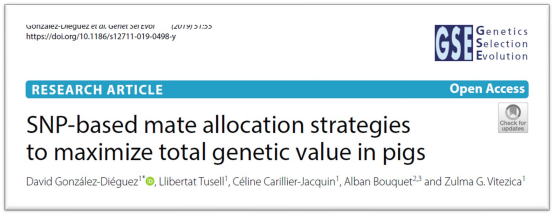


Example in pigs (within breed)

David González-Diéguez

France Genetic Porc
Age at 100 kg (AGE), Backfat depth (BD), Average piglet weight at birth (APWL)
39,353 SNPs

Trait	Boars	Sows	Genotyped animals	Number of records	Mean (SD)
AGE (days)	789	2179	2968	2968	149.03 (9.36)
BD (mm)	1007	2675	3682	3682	11.20 (1.68)
APWL (g)	1446	1226	2672	3297	1321.73 (213)

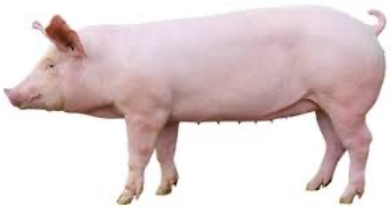

https://doi.org/10.1186/s12711-019-0498-y

Genetics Selection Evolution

RESEARCH ARTICLE Open Access

SNP-based mate allocation strategies to maximize total genetic value in pigs

David González-Diéguez[1*], Llibertat Tusell[1], Céline Carillier-Jacquin[1], Alban Bouquet[2,3] and Zulma G. Vitezica[1]



Landrace français

Example in pigs (within breed)

Estimation of variance components: $\sigma_A^2 \quad \sigma_D^2$

- Model GD : additive + dominance + genomic inbreeding

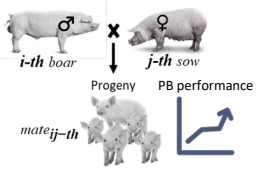
$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{F}b + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{e}$$

$\boldsymbol{F}$ is a vector of genomic inbreeding coefficients
b is the inbreeding depression parameter
$\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{G}\sigma_A^2)$, $\boldsymbol{G}$ built as in VanRaden (2008)
$\boldsymbol{v} \sim N(\boldsymbol{0}, \boldsymbol{D}\sigma_D^2)$, $\boldsymbol{D}$ built as in Vitezica *et al.* (2013)

- Model G : only additive + genomic inbreeding

remlf90 software (Misztal *et al.* 2012)

Within breed



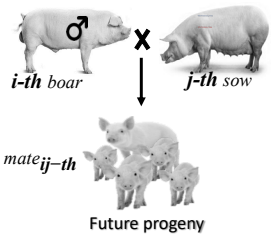
Estimation of additive and dominant SNP effects: $\hat{a}$ and $\hat{d}$

- BLUP-SNP model including dominance and genomic inbreeding

GS3 software (Legarra *et al.* 2011)

Example in pigs (within breed)

Prediction of expected progeny values (Toro and Varona 2010):



- Prediction of the total genetic values (g_{ij}) of the mating

$$\hat{g}_{ij} = \sum_k \left[P_{ijk}(CC)\hat{a}_k + P_{ijk}(CT)\hat{d}_k + P_{ijk}(TT)(-\hat{a}_k)\right]$$

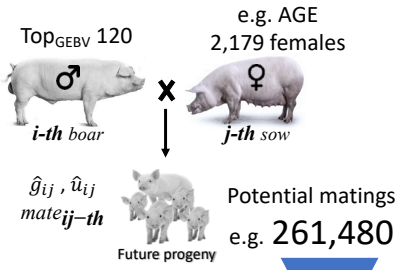
- Prediction of the breeding values (u_{ij}) of the progeny

$$\hat{u}_{ij} = \sum_k \left[P_{ijk}(CC)(2-2p_k)\hat{\alpha}_k + P_{ijk}(CT)(1-2p_k)\hat{\alpha}_k + P_{ijk}(TT)(-2p_k)\hat{\alpha}_k\right]$$

$$\hat{\alpha}_k = \hat{a}_k + \hat{d}_k(q_k - p_k)$$

Example in pigs (within breed)

Allocation of matings



Two mate allocation strategies:

(1) 600 matings selected on $\hat{u}_{ij} \rightarrow f_{optim}(\hat{u}_{ij})$

(2) 600 matings selected on $\hat{g}_{ij} \rightarrow f_{optim}(\hat{g}_{ij})$

Evaluation of expected genetic gains:

Additive genetic gain (Δu):

- $\Delta u = mean(\hat{u}_{600}) - mean(\hat{u}_{all\_matings})$

Total genetic superiority (Δg):

- $\Delta g = mean(\hat{g}_{600}) - mean(\hat{g}_{all\_matings})$

Optimization by linear programming

R package *lpsolve* (Berkelaar *et al.*, 2004)

Two constraints:

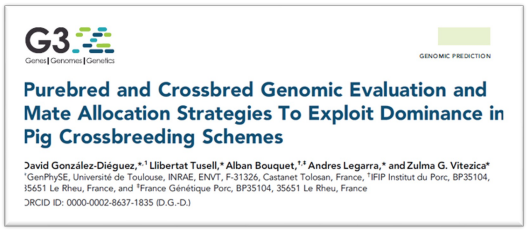
(1) each boar could be mated to up to 15 sows

(2) each sow could not be mated to more than one boar

Example in pigs (across breeds)

Is it possible to boost CB performance by implementing mate allocation in a two-way pig crossbreeding scheme in the long term?

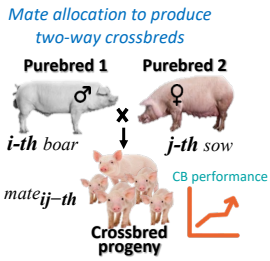
Simulation study (QMSim + Fortran program)
Maternal trait: litter size
Genome: 18 Chr 120 cM each


G3 Genes|Genomes|Genetics

GENOMIC PREDICTION

Purebred and Crossbred Genomic Evaluation and Mate Allocation Strategies To Exploit Dominance in Pig Crossbreeding Schemes

David González-Diéguez,*,† Llibertat Tusell,* Alban Bouquet,†,‡ Andres Legarra,* and Zulma G. Vitezica*
*GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France, †IFIP Institut du Porc, BP35104, 35651 Le Rheu, France, and ‡France Génétique Porc, BP35104, 35651 Le Rheu, France
ORCID ID: 0000-0002-8637-1835 (D.G.-D.)



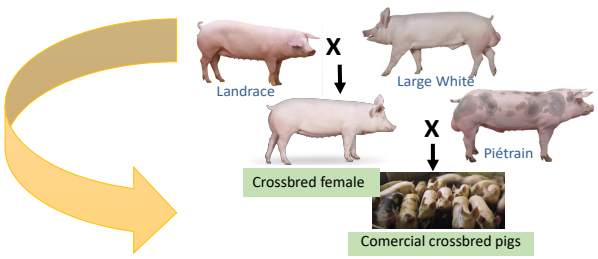
David González-Diéguez

Sargolzaei and Schenkel, 2009

Genetic improvement in pigs

- It uses selection and crossbreeding
- The breeding goal is to improve crossbred (CB) performance, while selection takes place in purebred (PB) animals based on PB performances
- Selection depends on the correlation between PB and CB performance (r_{PC})

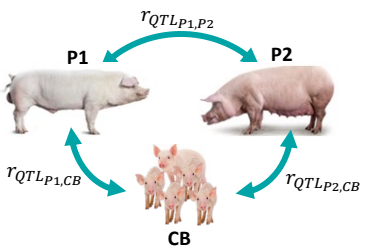
Selection may be suboptimal (GxE)
$r_{PC} < 1$ (~0.7)



Example in pigs (across breeds)

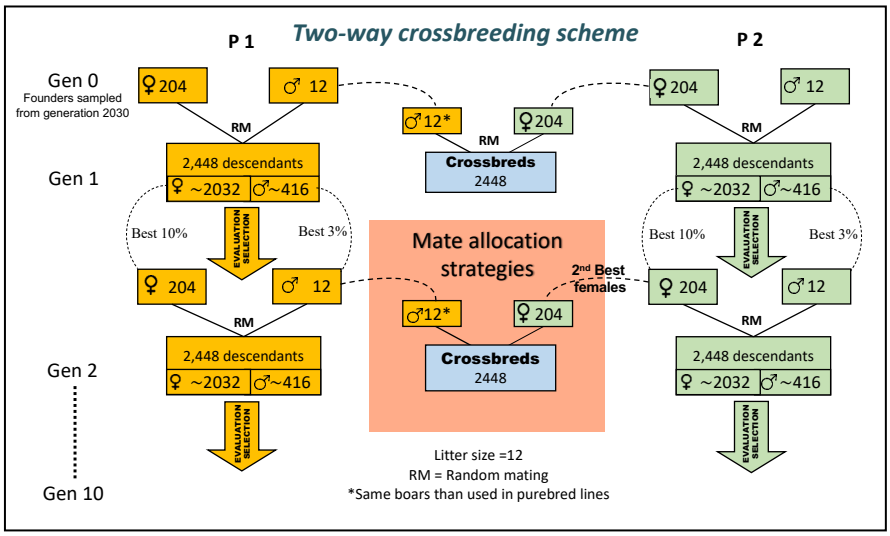
Simulation of heterosis and QTL effects

- Maternal trait: "e.g. Litter size" controlled by additive and dominant QTL action (2,500 QTLs)
- **Inbreeding depression** was assumed to be -1 piglet per 10% increase in genomic inbreeding in P1, P2 and CB
- Additive and dominance QTL effects were sampled from a MVN distribution with **correlation between the three populations to account for GxE and GxG**. Landrace and Yorkshire genetic variances were taken from Xiang *et al.* (2016)



Correlation between QTLs (r_{QTL}):
$r_{QTL_{P1,CB}} = r_{QTL_{P1,P2}} = r_{QTL_{P2,CB}} = 0.5$

Example in pigs (across breeds)



Mate allocation: results

- Mate allocation has a small added benefit within-breed and no benefit across-breed
- Selecting PB animals for CB performance using PB and CB data is a good strategy to exploit heterosis and improve crossbred performance, especially if the r_{PC} is low

Some conclusions

- We have a comprehensive theory
- We know how to properly define/estimate non-additive statistical effects
- Inbreeding/heterosis should be fit in the genetic evaluation model
- Fitting dominance and epistatic effects is interesting to correctly appraise genetic variances

Some conclusions

- Dominance and epistasis is not difficult with markers provided all animals ☺ (plants ☺) are genotyped
- In our experience, computational complexity is not an issue (models fit into computers), but convergence and accuracy are an issue (many parameters, little information)

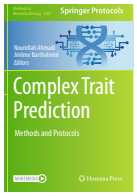
frontiers in Genetics

REVIEW
published: 06 March 2018
doi: 10.3389/fgene.2018.00078

Non-additive Effects in Genomic Selection

Luis Varona [1,2], Andres Legarra [3], Miguel A. Toro [4] and Zulma G. Vitezica [5]*

[1] Departamento de Anatomía, Embriología y Genética Animal, Universidad de Zaragoza, Zaragoza, Spain, [2] Instituto Agroalimentario de Aragón (IA2), Zaragoza, Spain, [3] Génétique Physiologie et Systèmes d'Elevage (GenPhySE), Institut National de la Recherche Agronomique de Toulouse, Castanet-Tolosan, France, [4] Departamento Producción Agraria, ETS Ingeniería Agronómica, Alimentaria y de Biosistemas, Universidad Politécnica de Madrid, Madrid, Spain, [5] Génétique Physiologie et Systèmes d'Elevage (GenPhySE), Université de Toulouse, Castanet-Tolosan, France



Chapter 8

Genomic Prediction Methods Accounting for Nonadditive Genetic Effects

Luis Varona, Andred Legarra, Miguel A. Toro, and Zulma G. Vitezica

Acknowledgments

Luis Varona
Miguel Angel Toro

Manolo Baselga
Juan Pablo Sanchez
Miguel Perez-Enciso

Daniela Laurenco
Ignacy Misztal

Ole Christensen

Toni Reverter
Fernanda Raidan

Andrés Legarra
Llibertat Tusell
Jean Michel Elsen
Alan Charcosset
Laurence Moreau

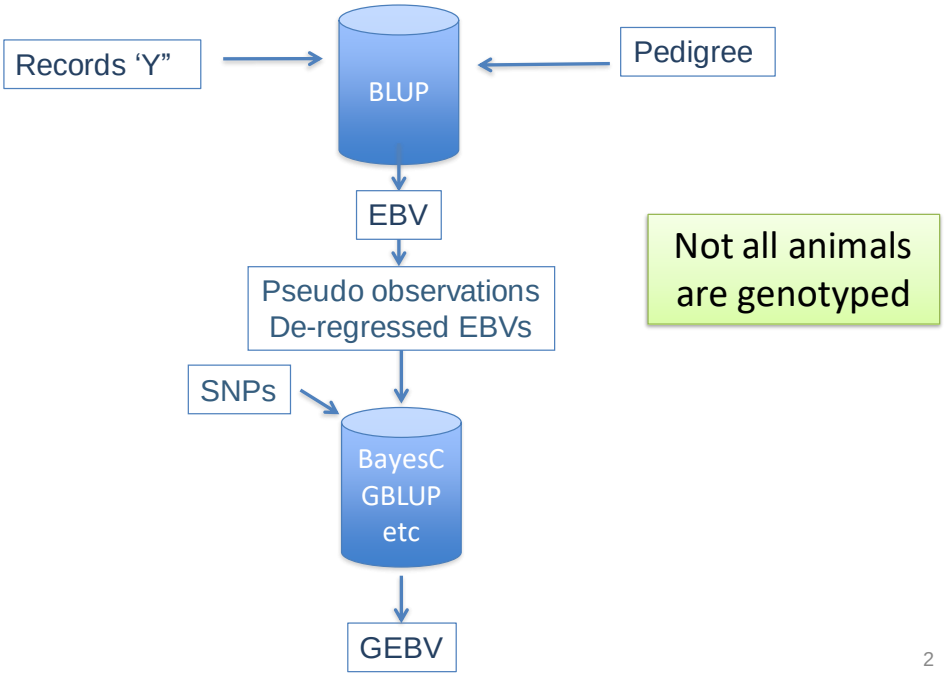
Johann Ertl
Eduardo Fernandez
Natalia Forneris
Carolina Garcia-Baccino
Tao Xiang
David Gonzalez-Dieguez

Christina Lehermeier
Simon Teyssedre

Thank you for your attention!

Matrix H and Single Step GBLUP

Multiple-step Genomic evaluation



Genomic evaluation

- Estimate effect of all SNPs in the genome:

$$\boldsymbol{y} = \boldsymbol{Za} + \boldsymbol{e}$$

$$\begin{pmatrix} 2 \\ 12 \\ 8 \\ 6.2 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 1 & -1 \\ 0 & -1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

Records

Genotypes

Marker effects

BayesA, BayesB, BLUP-SNP, etc

Genomic evaluation

- What about records?

$$\boldsymbol{y} = \boldsymbol{Za} + \boldsymbol{e}$$

$$\begin{pmatrix} 2 \\ 12 \\ 8 \\ 6.2 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 1 & -1 \\ 0 & -1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

Records

Genotypes

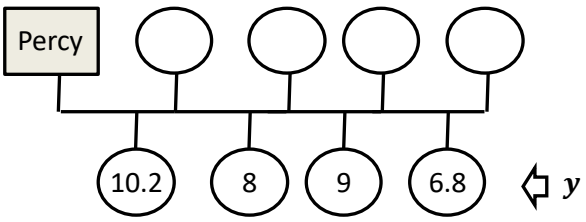
Marker effects

Genomic evaluation

- What about records?
- We genotype key animals (breeding males and maybe females)
  - They may *not* have phenotype on their own
  - They also have progenies who could have phenotype but could not have genotype
- "Project" family phenotypes on genotyped animals
  - Deregressed Proofs, DYD, etc.
  - Let's call this "DP"
- More easy said than done

What about records ?

- Typical in dairy cattle: the male is "assigned" the performance of the daughters
- Similar to a sire model

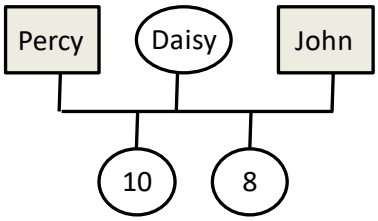


$$DP_{Percy} = \bar{y} = 8.5$$

- But to achieve more accuracy and to avoid selection bias, we need to correct for the dams' EBV and for the fixed effects
  - This is what we do in DEREGRESSION
- And corrections <u>contain errors</u> which pass on to deregressed proofs

What about records ?

- Assume that Daisy EBV is overestimated as 6.4 (true BV is 4)



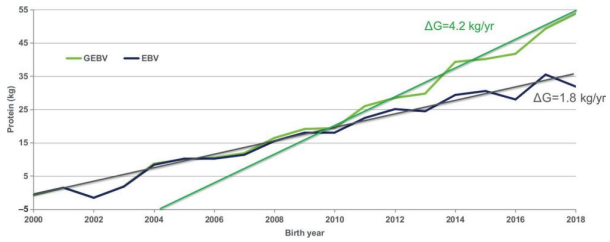
$$DP_{Percy} = 10 - 3.2 = 6.8$$
$$DP_{John} = 8 - 3.2 = 4.8$$

- Now both Percy and John are biased downwards !!
- Sometimes Daisy will be biased upwards and sometimes downwards
- Thus, the deregressed proofs of Percy and John will have a residual covariance
- This covariance is always ignored in practice

- The same problem exists when we correct by effects such as herd

Creeping bias

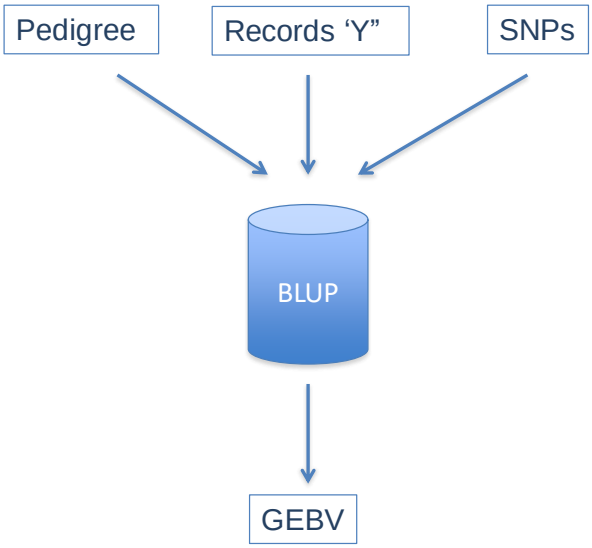
- BLUP can't consider selection that is not in the records
- thus, BLUP evaluations underestimate genomic selection trends



Mäntysaari, et al. (2020). JDS *103*(6), 5314-5326.

Single-Step Genomic Evaluation

Aguilar et al., 2010



Expand information

- We can do ONE evaluation if we "augment" information generating either
  - genotypes for all animals (SNP-BLUP)
  - **G** matrix for all animals (GBLUP)

  - genotypes of all animals will take lots of space (~100 M cows in US dairy cattle evaluation)
  - G matrix of all animals will be very cumbersome too
- ???

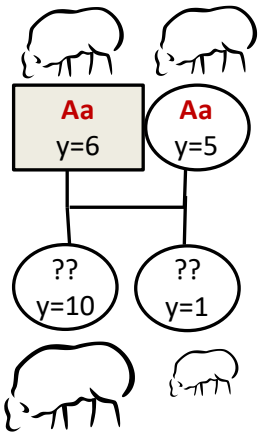
Expand information

- We can do ONE evaluation if we "augment" information generating either
  - genotypes for all animals (SNP-BLUP)
  - **G** matrix for all animals (GBLUP)

- Imputing algorithms (Beagle, Fimpute, AlphaImpute, etc.) are conceived to impute from low to high density
- For non-genotyped animals, we may "obtain" a point estimate of the genotype
- Why is this bad?

Problem with point estimates of genotypes

- Imagine a major gene



- Point estimate of genotype of the descendants: "Aa"
- Clearly, based on $\boldsymbol{y}$ there is Mendelian segregation where one descendant received "AA" and the other "aa"

- There is variation of true genotype around the point estimate of the genotype

- If we do not consider this variation we consider the offspring as identical twins

- Remember « genotype seen as a trait »?

Heritability of gene content

- If the genotype is accurate, "genotype seen as a trait" *z* is observed with no error
- *z* is transmitted from parents to offspring and there is no external influences
- *z* is additive (by definition)
- Heritability of *z* is 1 (!!!)

We can model gene content as a quantitative trait:

- $Cov(z_i, z_j) = A_{ij} 2pq$
- $\boldsymbol{z} = \mathbf{1}\mu + \boldsymbol{u} = \mathbf{1}(2p) + \boldsymbol{u}$
- $\mathrm{V}ar(u) = \boldsymbol{A}\sigma_u^2 = \boldsymbol{A}\sigma_z^2 = \boldsymbol{A}2pq$

Genotype prediction using BLUP for gene content (Gengler's method)

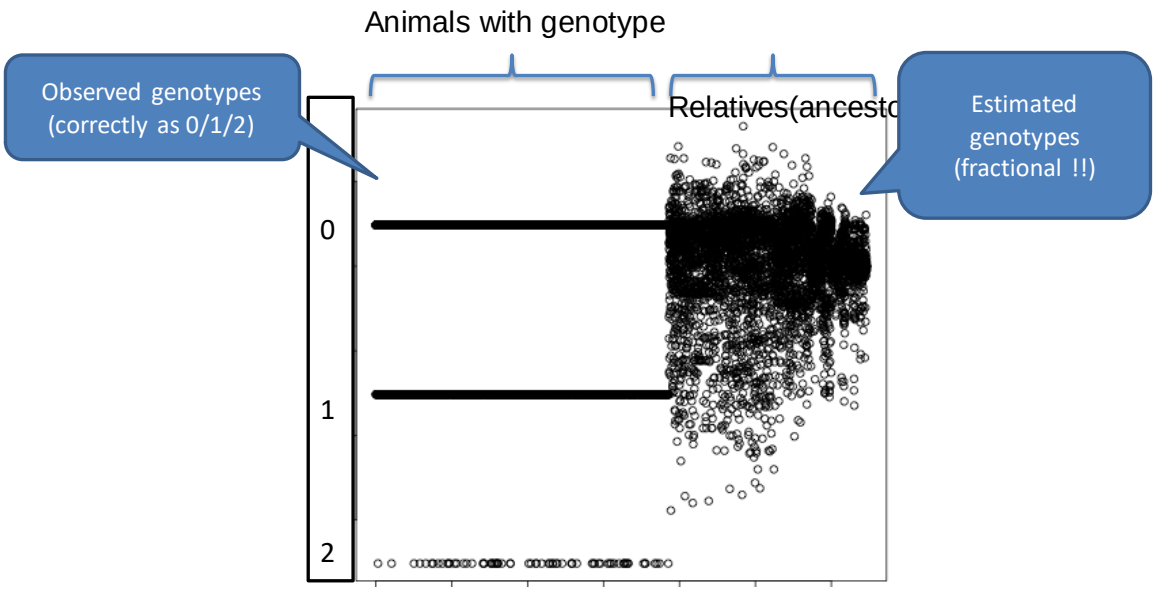
- Assuming $h^2 \approx 0.99$, use BLUP !!
  - $\boldsymbol{z} = \mathbf{1}\mu + \boldsymbol{W}\boldsymbol{u} + \boldsymbol{e}$

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{W} \\ \boldsymbol{W}'\boldsymbol{Z} & \boldsymbol{W}'\boldsymbol{W} + \boldsymbol{A}^{-1}\boldsymbol{\lambda} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\mu}} \\ \hat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{z} \\ \boldsymbol{W}'\boldsymbol{z} \end{pmatrix}$$

  - On exit, $\mathbf{1}\hat{\mu} + \hat{\boldsymbol{u}}$ are estimates of gene content for all animals

- Using Selection Index (which is BLUP without fixed effects)
  - if $\mathbf{z}$ has been centered using p in the base generation, then $\mu = 0$ and $\hat{\boldsymbol{u}}_2 = \boldsymbol{u}_2 = \boldsymbol{z}_2$ for genotyped animales, and
  - $\hat{\mathbf{z}}_1 = \hat{\mathbf{u}}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{z}_2$ is the *prediction* for non-genotyped animals

Example

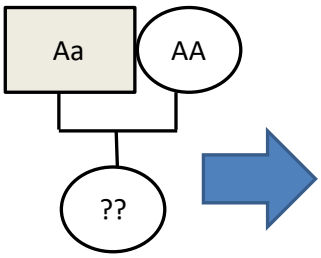


The prediction is not very good:
- it is fractional
- it has large error

Augmenting genotypes

- Gengler et al. (2007) conceived an algebraic way to deal with these point estimates
- Christensen & Lund (2010) showed how to take the error into account

- Genotype of descendants = half their parents + Mendelian sampling



AA with probability ½
Aa with probability ½

Prediction of Genotype = $\frac{3}{2}$"A" + $\frac{1}{2}$"a"

Variance(Genotype)=$\frac{1}{4}$"A" + $\frac{1}{4}$"a"

Missing data in « classic » contexts

Fill-in missing data: « data augmentation »

- Augmenting = adding genotypes
- But we need to account for the fact that these are « guesses »

- Expectation-Maximization, « data augmentation », « missing data theory »

Missing data in « fancy » contexts

Fill-in missing data: « imputation »

- « Hot deck imputation is **a method for handling missing data in which each missing value is replaced with an observed response from a "similar" unit**. ”
- Multiple imputation is **a general approach to the problem of missing data that is available in several commonly used statistical packages**. It aims to allow for the uncertainty about the missing data by creating several different plausible imputed data sets and appropriately combining results obtained from each of them

- What Expectation-Maximization or « missing data theory » does is an analytical »multiple hot deck imputation »

Single Step as a missing data problem

- We can see genotype as a missing data problem (Christensen & Lund, 2010)
- Use the prediction <u>and the distribution of the prediction</u> (if not the procedure does not work)

Let $$\mathbf{A}=\begin{bmatrix} \overbrace{\mathbf{A}_{11}}^{\textit{non genotyped}} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \underbrace{\mathbf{A}_{22}}_{\textit{genotyped}} \end{bmatrix}$$

Inferring genotypes

- There is Gengler's gene content prediction J. Dairy Sci. 91:1652
  - Linear approximation to the imputation problem
  - This method can be applied to any member of a pedigree
- Using <u>centered</u> gene content:

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

- Christensen and Lund realized that

$$Var(\hat{\boldsymbol{Z}}_1|\boldsymbol{Z}_2) = (\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})\boldsymbol{V}$$

where $\boldsymbol{V}$ contains (base population) $2p_k q_k$ in the diagonal

Inferring genotypes

- Instead of working with individual SNP effects, we will define
  - **u**=**Za**
  - i.e., the genetic value is the sum of SNP effects
  - We're not really interested in **a** themselves but in **u** (we know from GBLUP that we can jump from one to the other)
  - Moreover, we're interested in the distribution of **u**'s, so that we can compute their covariances and put them into the MME

Christensen & Lund key idea:

$$\boldsymbol{u} = \begin{pmatrix} \boldsymbol{u}_2 \\ \boldsymbol{u}_1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{Z}_2 \\ \boldsymbol{Z}_1 \end{pmatrix} \boldsymbol{a}$$

Breeding values

SNP effects

1= « non genotyped »
2= « genotyped »

Re-create GBLUP…

Chistensen & Lund use $Var(A) = E\big(Var(A|B)\big) + Var\big(E(A|B)\big)$ to consider the prediction of the genotype and its variance

$$\mathrm{Var}(\boldsymbol{u}) = \begin{pmatrix} \boldsymbol{Z}_2 \\ \widehat{\boldsymbol{Z}}_1 \end{pmatrix} Var(\boldsymbol{a})(\boldsymbol{Z}_2' \quad \widehat{\boldsymbol{Z}}_1) + \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & Var(\widehat{\boldsymbol{Z}}_1) \end{pmatrix} Var(\boldsymbol{a})$$

$E(\boldsymbol{Z}_1|\boldsymbol{Z}_2)$

$1/2\Sigma p_i q_i$

$Var(\boldsymbol{Z}_1|\boldsymbol{Z}_2)$

Using Gengler's results

Resulting in:

Covariances of all animals

Legarra et al. 2009; Aguilar et al., 2010; Christensen & Lund, 2010

$$Var\begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} = \mathbf{H} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{bmatrix} = \begin{bmatrix} \overbrace{\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}}^{\textit{non genotyped}} & \underbrace{\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}}_{\textit{genotyped}} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{bmatrix}$$

Let $\mathbf{A} = \begin{bmatrix} \overbrace{\mathbf{A}_{11}}^{\textit{non genotyped}} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$

Covariances of all animals

$$Var\begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix}=\mathbf{H}=\begin{bmatrix}\mathbf{H}_{11} & \mathbf{H}_{12}\\ \mathbf{H}_{21} & \mathbf{H}_{22}\end{bmatrix}=$$

$$\left[\begin{array}{c|c}\overbrace{\mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}}\overbrace{+\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\\ \hline \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G}\end{array}\right]$$

This is the variance of prediction of genotypes *from* genotyped *to* non-genotyped

This is the error in the prediction

The prediction « generates » a covariance

G comes from genotypes

$$Var\begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix}=\mathbf{H}=\begin{bmatrix}\mathbf{H}_{11} & \mathbf{H}_{12}\\ \mathbf{H}_{21} & \mathbf{H}_{22}\end{bmatrix}=$$

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

which after matrix algebra…

- Incredibly: $\mathbf{H}^{-1}$ is very simple:

$$\mathbf{H}^{-1}=\mathbf{A}^{-1}+\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}^{-1}-\mathbf{A}_{22}^{-1}\end{bmatrix}$$

…and avoiding « double counting »

Inverse of the regular pedigree relationship matrix

Correcting for genomic relationships…

- **Things would be simple if we had genomic relationships for everyone (Legarra et al., 2009)**
- Things would be simple if we could add genotypes for all animals (Christensen et al., 2010)

Overall modification

- Look at **A** as a « prior » relationship and to **G** as an « observed » relationship
  - **G** is observed for some individuals only, whose « a priori » relationship matrix was $\mathbf{A}_{22}$
- Try to construct a « posterior » relationship matrix

Joint distributions

Unconditional distribution of genetic values of Genotyped individuals

$$p(\mathbf{u}_2) = N(\mathbf{0}, \mathbf{G}) \text{ and}$$

After seeing their genotypes !

Conditional distribution of Non-Genotyped individuals

$$p(\mathbf{u}_1 | \mathbf{u}_2) = N\left(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2, \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)$$

Because they have no genotypes, this depends only on pedigree

$$p(\mathbf{u}_1, \mathbf{u}_2) = p(\mathbf{u}_2)\, p(\mathbf{u}_1 | \mathbf{u}_2)$$

Joint distribution

Joint distributions

$$p(\mathbf{u}_1, \mathbf{u}_2) = p(\mathbf{u}_1, | \mathbf{u}_2) p(\mathbf{u}_2)$$

$$= p(\mathbf{u}_1 | \mathbf{u}_2) p(\mathbf{u}_2)$$

prediction of non genotyped from genotyped

"Genomic" relationships

$$\propto \exp[-0.5(\mathbf{u}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2)'\mathbf{A}^{11}(\mathbf{u}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2)]\exp[-0.5\mathbf{u}_2'\mathbf{G}^{-1}\mathbf{u}_2]$$

$$= \exp\left(-0.5\begin{bmatrix}\mathbf{u}_1' & \mathbf{u}_2'\end{bmatrix}\begin{bmatrix}\mathbf{A}^{11} & -\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}^{11} & \mathbf{G}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{u}_1 \\ \mathbf{u}_2\end{bmatrix}\right)$$

$$= \exp\left(-0.5\begin{bmatrix}\mathbf{u}_1' & \mathbf{u}_2'\end{bmatrix}\begin{bmatrix}\mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{G}^{-1} + \mathbf{A}^{22} - \mathbf{A}_{22}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{u}_1 \\ \mathbf{u}_2\end{bmatrix}\right).$$

…for those inclined to algebra

Joint distributions

$$p(\mathbf{u}_2) = N(\mathbf{0}, \mathbf{G})$$

$$p(\mathbf{u}_1 | \mathbf{u}_2) = N\left(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2, \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)$$

Joint distributions

$$p(\mathbf{u}_2) = N(\mathbf{0}, \mathbf{G}) \quad \Rightarrow \quad Var(\mathbf{u}_2) = \mathbf{G}$$

$$p(\mathbf{u}_1 | \mathbf{u}_2) = N\left(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2, \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)$$

Joint distributions

$$p(\mathbf{u}_2) = N(\mathbf{0}, \mathbf{G}) \quad \Rightarrow \quad Var(\mathbf{u}_2) = \mathbf{G}$$

$$p(\mathbf{u}_1 | \mathbf{u}_2) = N\left(\underbrace{\mathbf{A}_{12}\mathbf{A}_{22}^{-1}}\mathbf{u}_2, \underbrace{\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}}\right)$$

$$Var(\mathbf{u}_1) = \overbrace{\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}} + \overbrace{\mathbf{A}_{12}\mathbf{A}_{22}^{-1}}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$$

because Var(**Xt**) = **X**Var(**t**)**X**'

Joint distributions

$$p(\mathbf{u}_2) = N(\mathbf{0}, \mathbf{G}) \quad \Rightarrow \quad Var(\mathbf{u}_2) = \mathbf{G}$$

$$p(\mathbf{u}_1 | \mathbf{u}_2) = N\left(\underbrace{\mathbf{A}_{12}\mathbf{A}_{22}^{-1}}\mathbf{u}_2, \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)$$

$$Cov(\mathbf{u}_1, \mathbf{u}_2) = \overbrace{\mathbf{A}_{12}\mathbf{A}_{22}^{-1}}\mathbf{G}$$

because Cov(**Xt,t**) = **X**Var(**t**)

Covariances of all animals

Legarra et al. 2009; Aguilar et al., 2010; Christensen & Lund, 2010

$$Var\begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix}=\mathbf{H}=\begin{bmatrix}\mathbf{H}_{11} & \mathbf{H}_{12}\\ \mathbf{H}_{21} & \mathbf{H}_{22}\end{bmatrix}=\begin{bmatrix}\overbrace{\mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}+\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}}^{\textit{non genotyped}} & \underbrace{\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}}_{\textit{genotyped}}\\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G}\end{bmatrix}$$

$$Var\begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix}=\mathbf{H}=\begin{bmatrix}\mathbf{H}_{11} & \mathbf{H}_{12}\\ \mathbf{H}_{21} & \mathbf{H}_{22}\end{bmatrix}=\begin{bmatrix}\mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}+\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G}\end{bmatrix}$$

- Incredibly: $\mathbf{H}^{-1}$ is very simple:

$$\mathbf{H}^{-1}=\mathbf{A}^{-1}+\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}^{-1}-\mathbf{A}_{22}^{-1}\end{bmatrix}$$

…and avoiding « double counting »

Inverse of the regular pedigree relationship matrix

Correcting for genomic relationships…

Understanding H matrix

- It is a projection of **G** matrix on the rest of individuals "so that" **G** matrix makes sense
  - e.g. parents of two animals related in **G** should be related in **A**
- It is a Bayesian updating of the pedigree relationship matrix based on new information from genotypes
- The approximation of multivariate normality is good because we have *many* markers

- Typically
  - $\mathbf{A}^{-1}$ in the millions but extremely sparse
  - **G** and $\mathbf{A}_{22}$ in the thousands
  - Leads to a very efficient method of genomic evaluation:
    - **Single Step GBLUP**

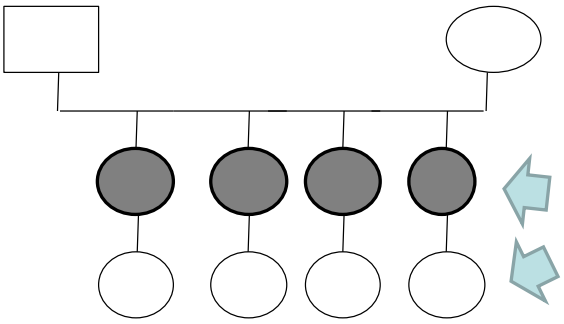
Understanding H matrix

- Still H it's an approximation: animals DO NOT have fractional genotypes
  - An optimal method would consider Mendelian inheritance, transmission and linkage disequilibrium
  - Which computationally and analytically is just too complicated

- My personal opinion is that **H** is good as far as we cover well key individuals at each generation
- For instance, if all AI males are genotyped
- But genotyping the last 2 years of animals and including the preceding 30 years of pedigreed animals in H might not be a good idea
-

Examples on H matrix

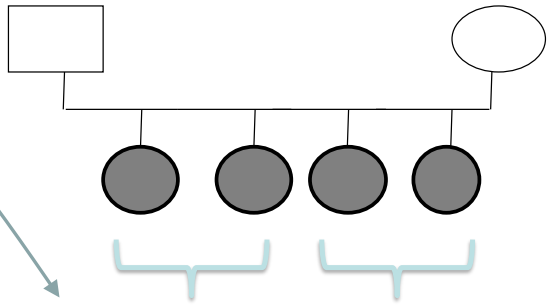
- Consider 4 full-sibs with one progeny each



- With pedigree, sibs are related by 0.5
- their offspring are cousins with a relationship of 0.125

- The 0.5 assumes infinite unlinked loci, with actual genomes relationship varies: 0.5 ± 0.05

- Pedigree; grey is genotyped



G=

1.0	0.7	0.3	0.3
0.7	1.0	0.3	0.3
0.3	0.3	1.0	0.7
0.3	0.3	0.7	1.0

More related than average

More related than average

Less related than average

How is this included in **A**?

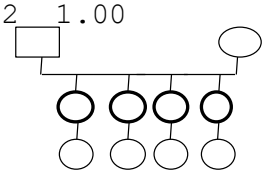
Classical A (pedigree)

1.00	0.00	0.50	0.50	0.50	0.50	0.25	0.25	0.25	0.25
0.00	1.00	0.50	0.50	0.50	0.50	0.25	0.25	0.25	0.25
0.50	0.50	1.00	0.50	0.50	0.50	0.50	0.25	0.25	0.25
0.50	0.50	0.50	1.00	0.50	0.50	0.25	0.50	0.25	0.25
0.50	0.50	0.50	0.50	1.00	0.50	0.25	0.25	0.50	0.25
0.50	0.50	0.50	0.50	0.50	1.00	0.25	0.25	0.25	0.50
0.25	0.25	0.50	0.25	0.25	0.25	1.00	0.12	0.12	0.12
0.25	0.25	0.25	0.50	0.25	0.25	0.12	1.00	0.12	0.12
0.25	0.25	0.25	0.25	0.50	0.25	0.12	0.12	1.00	0.12
0.25	0.25	0.25	0.25	0.25	0.50	0.12	0.12	0.12	1.00

Full-sibs is 0.50

Cousins is 0.125

Uncle-nephew is 0.25

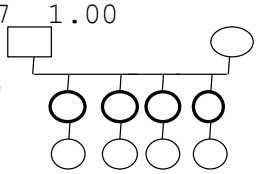


H (pedigree + markers)

1.00	0.00	0.50	0.50	0.50	0.50	0.25	0.25	0.25	0.25
0.00	1.00	0.50	0.50	0.50	0.50	0.25	0.25	0.25	0.25
0.50	0.50	1.00	0.70	0.30	0.30	0.50	0.35	0.15	0.15
0.50	0.50	0.70	1.00	0.30	0.30	0.35	0.50	0.15	0.15
0.50	0.50	0.30	0.30	1.00	0.70	0.15	0.15	0.50	0.35
0.50	0.50	0.30	0.30	0.70	1.00	0.15	0.15	0.35	0.50
0.25	0.25	0.50	0.35	0.15	0.15	1.00	0.17	0.07	0.07
0.25	0.25	0.35	0.50	0.15	0.15	0.17	1.00	0.07	0.07
0.25	0.25	0.15	0.15	0.50	0.35	0.07	0.07	1.00	0.17
0.25	0.25	0.15	0.15	0.35	0.50	0.07	0.07	0.17	1.00

Full-sibs is 0.70 – 0.30

Uncle-nephew is 0.35-0.15

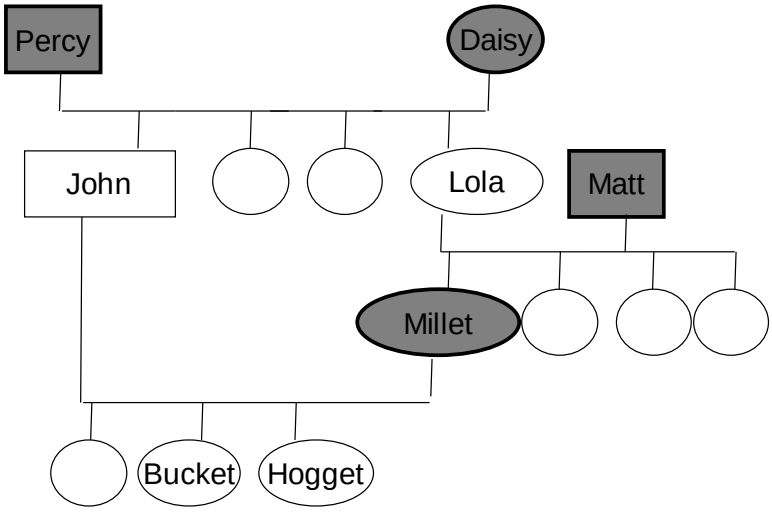


Cousins is 0.17 - 0.07

We have extended genomic relationships to all the pedigree

More complex example

- Pedigree; grey is genotyped

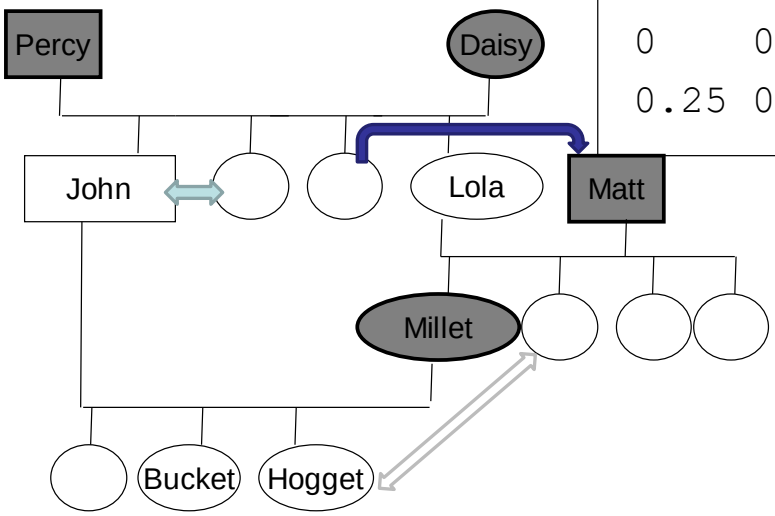


More complex example

- Before genotyping

$\mathbf{A}_{22}$=

1.00	0	0	0.25
0	1	0	0.25
0	0	1	0.50
0.25	0.25	0.50	1.00

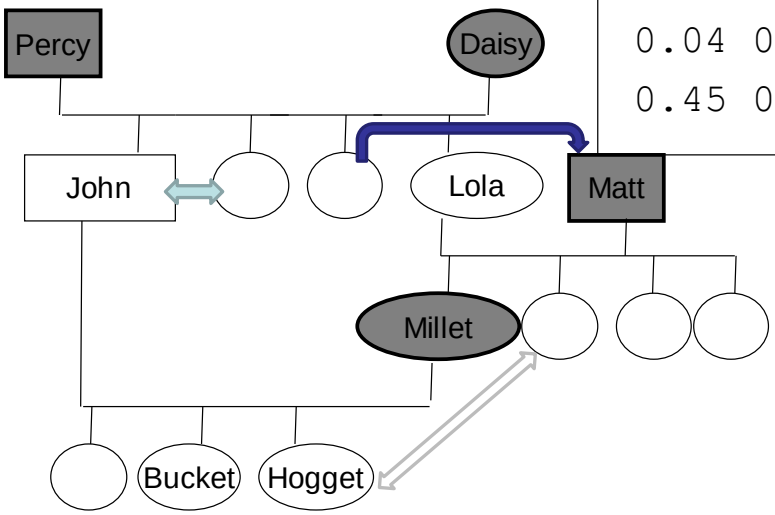


More complex example

- After genotyping

G=

1.13	0.08	0.04	0.45
0.08	0.91	0.14	0.32
0.04	0.14	1.12	0.62
0.45	0.32	0.62	1.10



Classical A (pedigree)

1.00	0.00	0.00	0.50	0.50	0.50	0.5	0.25	0.25	0.25	0.25	0.38	0.38	0.38
0.00	1.00	0.00	0.50	0.50	0.50	0.5	0.25	0.25	0.25	0.25	0.38	0.38	0.38
0.00	0.00	1.00	0.00	0.00	0.00	0.0	0.50	0.50	0.50	0.50	0.25	0.25	0.25
0.50	0.50	0.00	1.00	0.50	0.50	0.5	0.25	0.25	0.25	0.25	0.62	0.62	0.62
0.50	0.50	0.00	0.50	1.00	0.50	0.5	0.25	0.25	0.25	0.25	0.38	0.38	0.38
0.50	0.50	0.00	0.50	0.50	1.00	0.5	0.25	0.25	0.25	0.25	0.38	0.38	0.38
0.50	0.50	0.00	0.50	0.50	0.50	1.0	0.50	0.50	0.50	0.50	0.50	0.50	0.50
0.25	0.25	0.50	0.25	0.25	0.25	0.5	1.00	0.50	0.50	0.50	0.62	0.62	0.62
0.25	0.25	0.50	0.25	0.25	0.25	0.5	0.50	1.00	0.50	0.50	0.38	0.38	0.38
0.25	0.25	0.50	0.25	0.25	0.25	0.5	0.50	0.50	1.00	0.50	0.38	0.38	0.38
0.25	0.25	0.50	0.25	0.25	0.25	0.5	0.50	0.50	0.50	1.00	0.38	0.38	0.38
0.38	0.38	0.25	0.62	0.38	0.38	0.5	0.62	0.38	0.38	0.38	1.12	0.62	0.62
0.38	0.38	0.25	0.62	0.38	0.38	0.5	0.62	0.38	0.38	0.38	0.62	1.12	0.62
0.38	0.38	0.25	0.62	0.38	0.38	0.5	0.62	0.38	0.38	0.38	0.62	0.62	1.12

Full-sibs is 0.50

Uncle-nephew is 0.38

Unrelated is 0

H (pedigree + markers)

1.13	0.08	0.04	0.60	0.60	0.60	0.65	0.45	0.34	0.34	0.34	0.52	0.52	0.52
0.08	0.91	0.14	0.50	0.50	0.50	0.50	0.32	0.32	0.32	0.32	0.41	0.41	0.41
0.04	0.14	1.12	0.09	0.09	0.09	0.09	0.62	0.61	0.61	0.61	0.35	0.35	0.35
0.60	0.50	0.09	1.05	0.55	0.55	0.58	0.38	0.33	0.33	0.33	0.72	0.72	0.72
0.60	0.50	0.09	0.55	1.05	0.55	0.58	0.38	0.33	0.33	0.33	0.47	0.47	0.47
0.60	0.50	0.09	0.55	0.55	1.05	0.58	0.38	0.33	0.33	0.33	0.47	0.47	0.47
0.65	0.50	0.09	0.58	0.58	0.58	1.09	0.62	0.59	0.59	0.59	0.60	0.60	0.60
0.45	0.32	0.62	0.38	0.38	0.38	0.62	1.10	0.62	0.62	0.62	0.74	0.74	0.74
0.34	0.32	0.61	0.33	0.33	0.33	0.59	0.62	1.10	0.60	0.60	0.48	0.48	0.48
0.34	0.32	0.61	0.33	0.33	0.33	0.59	0.62	0.60	1.10	0.60	0.48	0.48	0.48
0.34	0.32	0.61	0.33	0.33	0.33	0.59	0.62	0.60	0.60	1.10	0.48	0.48	0.48
0.52	0.41	0.35	0.72	0.47	0.47	0.60	0.74	0.48	0.48	0.48	1.23	0.73	0.73
0.52	0.41	0.35	0.72	0.47	0.47	0.60	0.74	0.48	0.48	0.48	0.73	1.23	0.73
0.52	0.41	0.35	0.72	0.47	0.47	0.60	0.74	0.48	0.48	0.48	0.73	0.73	1.23

Full-sibs is 0.55

Uncle-nephew is 0.48

"Unrelated" is 0.14 ⇨ Because pedigree founders are related in **G**

Some properties of **H**

- Semi-positive definite <u>always</u>
- Positive definite & invertible <u>if & only if</u> **G** is invertible
- If everyone is genotyped, Single Step is GBLUP
- If no one is genotyped, Single Step is BLUP
- In practice, if **G** is too different from $\mathbf{A}_{22}$, this gives lots of numerical problems
  - (wrong pedigree or genotyping)
  - very poor « compatibility »

H matrix

- H is then a relationship matrix constructed with markers and pedigree
- But Henderson taught us how to use relationship matrices of any kind

Single step GBLUP

Single Step = Your regular BLUP with small modifications

W: incidence matrix of animals on data

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}\mathbf{R}^{-1}\mathbf{W} + \boxed{\mathbf{H}^{-1}}\sigma_u^{-2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

G

This **G** could be *any* matrix describing « genomic » covariances of breeding values;

it does not restrict to VanRaden's (2008) GBLUP

A: pedigree relationship matrix

$\mathbf{A}_{22}$: pedigree matrix among genotyped individuals

Single step GBLUP

- So the Single Step GBLUP is like regular BLUP changing one small submatrix !!!
- It is almost too simple to be true…

Single Step GBLUP

- Easy modification to a general purpose BLUP software
  - Only changes: addition of $\boldsymbol{G}^{-1}$ and $\boldsymbol{A}_{22}^{-1}$
  - Matrices $\boldsymbol{G}^{-1}$ and $\boldsymbol{A}_{22}^{-1}$ can be computed with external tools
- <u>Can fit any model</u> (probit, GxE,…)
- Simple extraction of SNP effects for indirect prediction or (multimarker) GWAS:

$$\hat{\boldsymbol{a}} = \frac{\boldsymbol{Z}'\boldsymbol{G}^{-1}\hat{\boldsymbol{u}}_2}{2\sum pq}$$

- Avoids selection bias due to genomic preselection

  (Patry & Ducrocq, 2011)

Single Step GBLUP

- What models have we fit so far in SSGBLUP?
  - Multiple traits (up to 18 so far)
  - Multiple trait + correlated genetic maternal effects (beef cattle)
  - Random regressions (lactation curves)
  - Threshold (probit) models
  - Horse rankings (Thurstonian model)
- *Anything* that was fit in BLUP can be fit in SSGBLUP, changing **A** to **H**

Details in ssGBLUP

Details in SSGBLUP

- Storage
- Inbreeding
- G is not invertible (« blending »)
- G might not explain all genetic variance (« blending »)
- Compatibility of G and A22
  - Assumption p($\mathbf{u}_2$)=N(**0**,**G**)
    - If there is selection, mean is not **0** (« tuning » solves it: see Vitezica later)
  - Same genetic variance in genotyped and ungenotyped animals
- Large data
- Unknown parent groups
  - Need to modify **H** to include them (Misztal et al., 2013)
  - Metafounders
- Crosses
- Computation
  - APY
  - Sherman-Woodbury
  - « hybrid » model

Details in SSGBLUP

- Storage
- Inbreeding
- G is not invertible (« blending »)
- G might not explain all genetic variance (« blending »)
- Compatibility of G and A22
  - Assumption p($\mathbf{u}_2$)=N($\mathbf{0}$,$\mathbf{G}$)
    - If there is selection, mean is not **0** (« tuning » solves it: see Vitezica later)
  - Same genetic variance in genotyped and ungenotyped animals
- Unknown parent groups
  - Need to modify **H** to include them (Misztal et al., 2013)
  - Metafounders
- Crosses
- Computation
  - APY
  - Sherman-Woodbury
  - « hybrid » model

Storage

$$\boldsymbol{H}^{-1} = \boldsymbol{A}^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{G}^{-1} - \boldsymbol{A}_{22}^{-1} \end{pmatrix}$$

$\boldsymbol{A}^{-1}$ is very sparse (9 elements /animal)

$\boldsymbol{G}^{-1} - \boldsymbol{A}_{22}^{-1}$ is very dense (number of genotyped animals^2)

Efficient storage and handling using hash/ija/yams

When $\boldsymbol{G}^{-1} - \boldsymbol{A}_{22}^{-1}$ is very big, use APY or similar methods

Manech Tete Rouse sheep:
- 3000 animals (rams) genotyped
- 500,000 animals pedigree.
- $\boldsymbol{A}^{-1}$ ~ 36 Mb RAM
- $\boldsymbol{H}^{-1}$ ~ 108 Mb

Angus beef cattle:
- 500,000 animals genotyped
- 11M animals pedigree.
- $\boldsymbol{A}^{-1}$ ~ 800 Mb RAM
- $\boldsymbol{H}^{-1}$ has 350 $\times 10^9$elements ~ 2800 Gb !

Details in SSGBLUP

- Storage
- **Inbreeding**
- G is not invertible (« blending »)
- G might not explain all genetic variance (« blending »)
- Compatibility of G and A22
  - Assumption p($\mathbf{u}_2$)=N(**0**,**G**)
    - If there is selection, mean is not **0** (« tuning » solves it: see Vitezica later)
  - Same genetic variance in genotyped and ungenotyped animals
- Unknown parent groups
  - Need to modify **H** to include them (Misztal et al., 2013)
  - Metafounders
- Crosses
- Computation
  - APY
  - Sherman-Woodbury
  - « hybrid » model

Inbreeding

- Inbreeding F_i is useful to:
  - Monitor genetic diversity
  - Obtain accuracies as $acc_i = \sqrt{1 - \frac{PEV_i}{(1+F_i)}}$
- Obtaining inbreeding in $\boldsymbol{A}$ is easy $\boldsymbol{F}_{\boldsymbol{A}i} = \boldsymbol{A}_{ii} - 1$ (e.g. Meuwissen and Luo 1992)
- Obtaining inbreeding in $\boldsymbol{G}$ is easy $\boldsymbol{F}_{\boldsymbol{G}i} = \boldsymbol{G}_{ii} - 1 = \frac{\boldsymbol{z}_i \boldsymbol{z}_i'}{\mathbf{2}\sum p_j q_j} - 1$
- Obtaining inbreeding in $\boldsymbol{H}$ is very complicated !!

Computing H inbreeding

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

- We only need for individual *i* :

$$h_{ii} = a_{ii} + \boldsymbol{a}_{i,2}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\boldsymbol{a}_{2,i} = a_{ii} + \boldsymbol{a}'_{2,i}\mathbf{M}\boldsymbol{a}_{2,i}$$

Or

$$F_{H_i} = F_{A_i} + \boldsymbol{a}'_{2,i}\mathbf{M}\boldsymbol{a}_{2,i} = F_{A_i} + c_i$$

- $\boldsymbol{a}_{2,i}$: relationships of *i* with all the genotyped individuals
- All the difficulty is computing correction c_i

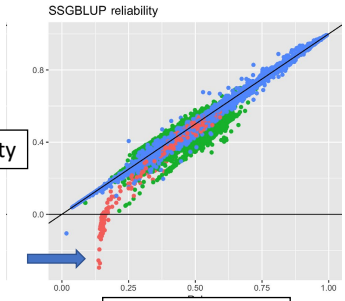
Inbreeding

Short communication: **Methods to compute genomic inbreeding for ungenotyped individuals**

A. Legarra,[1*] I. Aguilar,[2] and J. J. Colleau[3]

For ungenotyped animals, $H_{ii} = A_{ii} + c_i$ and $F_{Hi} = F_{Ai} + c_i$, where F_{Ai} is the pedigree inbreeding coefficient. The correction term c_i is equal to the quadratic $c_i = \mathbf{a}_{i,2}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{a}_{2,i} - \mathbf{a}_{i,2}\mathbf{A}_{22}^{-1}\mathbf{a}_{2,i} = \mathbf{a}'_{2,i}\mathbf{M}\mathbf{a}_{2,i}$, where $\mathbf{a}'_{i,2} = \mathbf{a}_{2,i}$ (the ith column of matrix $\mathbf{A}_{21}$) is the vector of relationships between ungenotyped individual i with all the genotyped individuals and $\mathbf{M} = \mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}$.

- 3 methods (all rather technical)
- Not doing things well leads to unexpected results, i.e. negative accuracies



Incorrect reliability

Correct reliability

Effects of ignoring inbreeding in model-based accuracy for BLUP and SSGBLUP

Ignacio Aguilar[1] | Eduardo N. Fernandez[2] | Agustin Blasco[3] | Olga Ravagnolo[1] | Andres Legarra[4]

Obtaining overall measures of diversity

- Optimal Contribution
- If you select bulls in list **x** and cows in list **y**, the expected future inbreeding is $\boldsymbol{y}'\boldsymbol{H}\boldsymbol{x} = \boldsymbol{y}'(\boldsymbol{H}\boldsymbol{x})$
- Optimal contribution methods optimize the lists to minimize inbreeding while keeping genetic progress (e.g. choosing two cousins instead of two sibs)
- For the pedigree information, obtaining $\boldsymbol{A}\boldsymbol{x}$ is very easy using the algorithm by Colleau (2002)
- Modification of the algorithm to obtain $\boldsymbol{H}\boldsymbol{x}$

Obtaining overall measures of diversity

- Global measures of diversity (e.g. average relationship of all young bulls) can be obtained as $\boldsymbol{x}'\boldsymbol{H}\boldsymbol{x} = \boldsymbol{x}'(\boldsymbol{H}\boldsymbol{x})$
- Obtaining $\boldsymbol{A}\boldsymbol{x}$ is very easy using the algorithm by Colleau
- Modification of the algorithm to obtain $\boldsymbol{H}\boldsymbol{x}$

1. Compute $\mathbf{z} = \mathbf{A}\mathbf{x}$ using [4],
2. Compute $\mathbf{y}_2 = \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{z}_2 = \mathbf{G}\left(\mathbf{A}_{22}^{-1}\mathbf{z}_2\right)$,
3. Compute $\mathbf{d}_2 = \mathbf{y}_2 - \mathbf{z}_2$,
4. Compute $\mathbf{d}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{d}_2$,
5. Compute $\mathbf{y}_1 = \mathbf{z}_1 + \mathbf{d}_1$. This is the final step.

A fast indirect method to compute functions of genomic relationships concerning genotyped and ungenotyped individuals, for diversity management

Jean-Jacques Colleau[1], Isabelle Palhière[2], Silvia T. Rodríguez-Ramilo[2] and Andres Legarra[2*]

Details in SSGBLUP

- Storage
- Inbreeding
- G is not invertible (« blending »)
- G might not explain all genetic variance (« blending »)
- Compatibility of G and A22
  - Assumption $p(\mathbf{u}_2)=N(\mathbf{0},\mathbf{G})$
    - If there is selection, mean is not **0** (« tuning » solves it: see Vitezica later)
  - Same genetic variance in genotyped and ungenotyped animals
- Unknown parent groups
  - Need to modify **H** to include them (Misztal et al., 2013)
  - Metafounders
- Crosses
- Computation
  - APY
  - Sherman-Woodbury
  - « hybrid » model

Blending and compatibility

- These are two different things
- Many people don't understand this
- "compatibility" or "tuning" tries to put **G** and **A** in the same scale
- "blending" :
  - assigns part of the genetic variance to pedigree – not markers
  - at the same time used to have an invertible **G**.

- we have seen this in the GBLUP part
- I will explain now why this might be important

Details in SSGBLUP

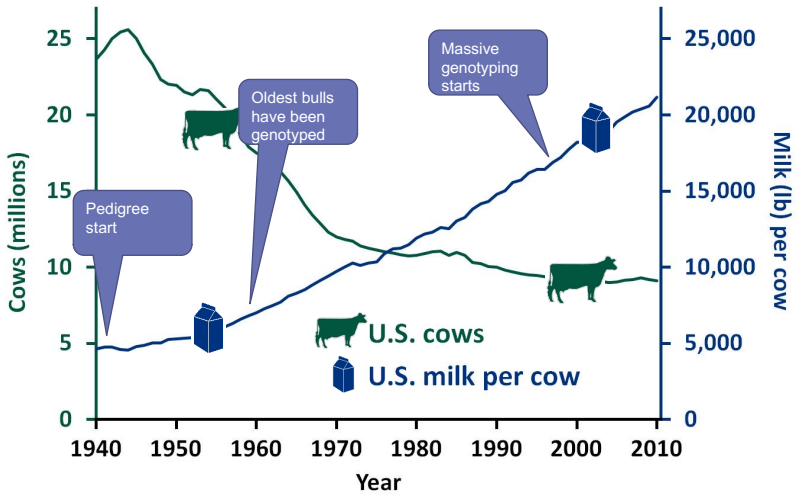
- Storage
- Inbreeding
- G is not invertible (« blending »)
- G might not explain all genetic variance (« blending »)
- Compatibility of G and A22
  - Assumption $p(\mathbf{u}_2)=N(\mathbf{0},\mathbf{G})$
    - If there is selection, mean is not **0** (« tuning » solves it: see Vitezica later)
  - Same genetic variance in genotyped and ungenotyped animals
- Unknown parent groups
  - Need to modify **H** to include them (Misztal et al., 2013)
  - Metafounders
- Crosses

- Computation
  - APY
  - Sherman-Woodbury
  - « hybrid » model

Compatibility of marker and pedigree relationships

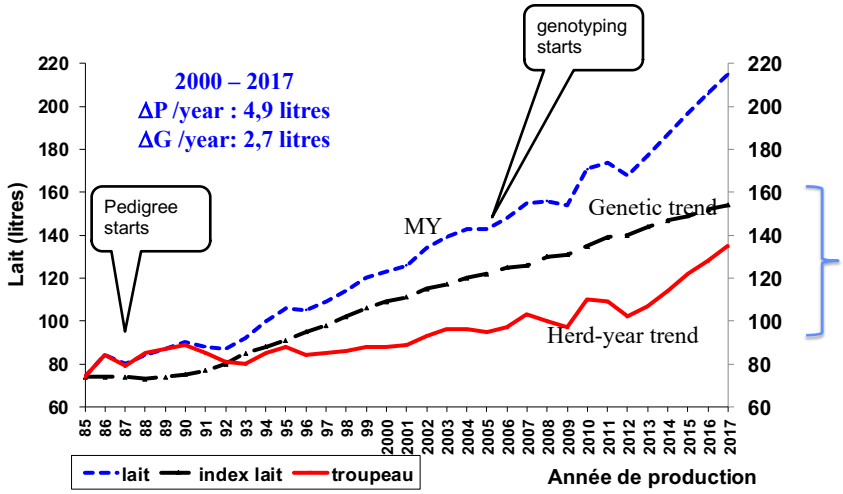
- Populations evolve with time, but genotypes came years after pedigree started
- Underlying hypothesis false:
  - Christensen & Lund (base allelic frequencies known)
  - Legarra et al. (average genetic value does not change)
- Genomic Predictions may be shifted from Pedigree Predictions
  - and make them not directly comparable

U.S. dairy population and milk yield



Milk Yield Genetic and Herd-year trend

MANECH TETE NOIRE



This ΔG from "pedigree start" to "genotype starts" needs to be considered

Compatibility of marker and pedigree relationships

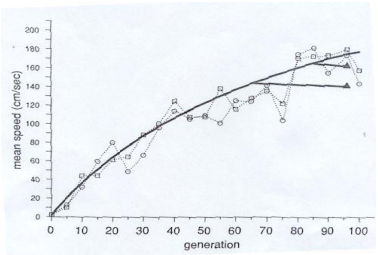
- The population for which $\text{average}(\boldsymbol{u}) = 0$ and for which the genetic variance is defined is called the *genetic base*
  - *Founders of the pedigree in classical* ***A***
  - *Whole set of genotyped animals in most typical* ***G***

- Typically, genotyped animals come *after* pedigree starts
  - e.g. Lacaune sheep pedigree go back to 1960 but genotypes start in 1995
- Drift (and selection) causes :
  - Average genetic values "drift" (in particular in small populations)
  - Genetic variance reduces

Reduction of genetic variance

Long-term selection experiments (Weber, 1996)

Two populations of Drosophila selected for performance in a wind tunnel with effective sizes 500-1000 and selected proportion of 4.5%.

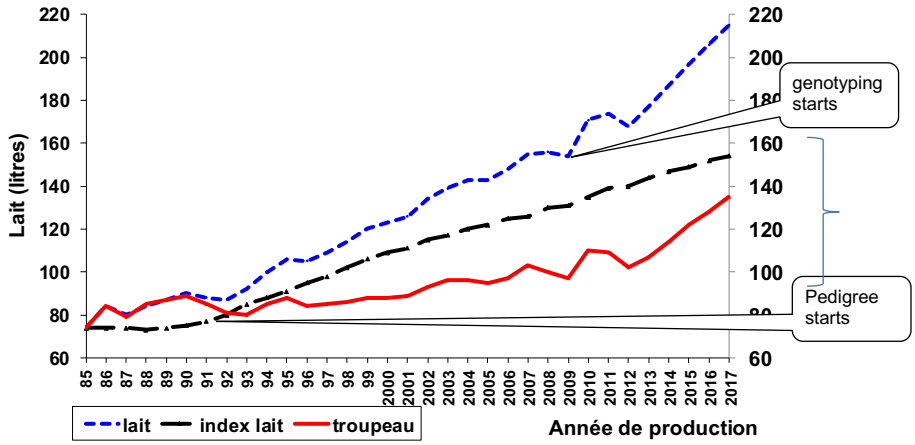


Cut data

- For practical purposes, you only need a few years of data
- Simplest thing: cut old data and pedigree
- Then there is no problem of selection and $p_{base} \approx p_{current}$
- Lourenco (2014) did this with good results
- Many breeds are reluctant because they feel that they loose information

Force G to be similar to A

This ΔG from "pedigree start" to "genotype starts" needs to be considered



$$p(\boldsymbol{u}_2) = N(\mathbf{0}, \boldsymbol{G}\sigma_u^2)$$

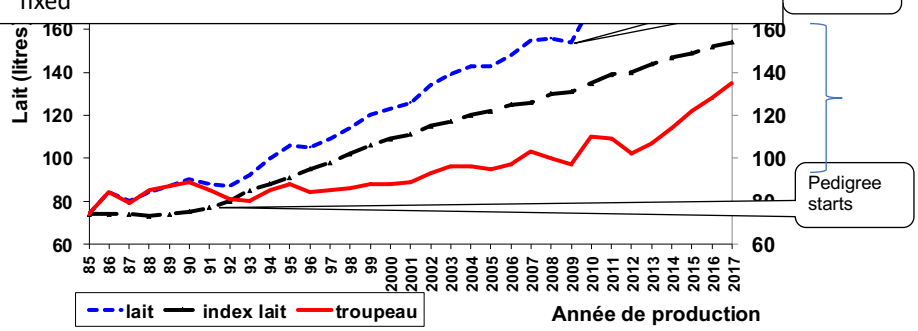
Setting both to 0 does not make sense

$$p(\boldsymbol{u}) = N(\mathbf{0}, \boldsymbol{A}\sigma_u^2)$$

Force G to be similar to A

Vitezica et al. 2011 included the ΔG explicitely as un unknown μ

- μ is random because $\mu = \Delta G = \overline{\boldsymbol{u}_2} = \frac{1}{n}\mathbf{1}'\boldsymbol{u}_2$
- μ has variance $\mathrm{Var}(\Delta G) = \alpha = Var(\overline{u_2}) = \frac{1}{n_2^2}\mathbf{1}'\boldsymbol{A}_{22}\mathbf{1}$ for typical **G**
- Fernando et al. (2014, 2016) method of J- coefficients consider μ as fixed



$$p(\boldsymbol{u}_2) = N(\mathbf{1}\mu, \boldsymbol{G}\sigma_u^2)$$

Setting both to 0 does not make sense

$$p(\boldsymbol{u}) = N(\mathbf{0}, \boldsymbol{A}\sigma_u^2)$$

Force G to be similar to A

- You can include explicitly:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} & \mathbf{0} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{H}^{-1}\lambda & -\mathbf{H}^{-1}\mathbf{Q}\lambda \\ \mathbf{0} & -\mathbf{Q}'\mathbf{H}^{-1}\lambda & \mathbf{Q}'\mathbf{H}^{-1}\mathbf{Q}\lambda+\alpha^{-1}\lambda \end{bmatrix} \times \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \\ \mu \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ 0 \end{bmatrix}.$$

- Or implicitly (equivalent model)

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{H}^{\dagger\,-1}\lambda \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix},$$

where

$$\mathbf{H}^{\dagger\,-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22}+(\mathbf{G}+\mathbf{1}\mathbf{1}'\alpha)^{-1}-\mathbf{A}_{22}^{-1} \end{bmatrix}.$$

Force G to be similar to A

- The method has an interesting genetic interpretation
- Using $\boldsymbol{G} \leftarrow \boldsymbol{G} + \mathbf{1}\mathbf{1}'\alpha$ forces **G** to yield same "average relationship" than $\boldsymbol{A}_{22}$

- But we forgot something…
- There is reduction in the genetic variance
- This reduction is contained in the inbreeding coefficients
- Thus, we should have $\mathrm{diag}(\boldsymbol{G}) \approx diag(\boldsymbol{A}_{22})$

Force G to be similar to A

- Vitezica et al. (2011) and Christensen et al. (2012) provided an unbiased method that forces the same genetic base across **G** and **A** :

$$\boldsymbol{G}^* = a + b\boldsymbol{G}$$

  - a accounts for old relationships among non genotyped ancestors
  - b accounts for reduction in the genetic variance

$$a + b\,\overline{\boldsymbol{G}} = \overline{\boldsymbol{A}}_{22}$$

$$a + b\,\overline{diag(\boldsymbol{G})} = \overline{\left(diag(\boldsymbol{A_{22}})\right)}$$

Does actually G resemble A?

- If pedigree is good and genotyping is good, yes
- Usually
  - $Cor\left(A_{22_{ij}}, G_{ij}\right) \approx 0.8$
  - $Cor\left(F_{pedigree_i}, F_{genomic_i}\right) \approx 0.5$
- Useful for quality control

Does actually G resemble A?

Differences between genomic-based and pedigree-based relationships in a chicken population, as a function of quality control and pedigree links among individuals

H. Wang[1], I. Misztal[2] & A. Legarra[3]

Table 2 Statistics for coefficient differences between genomic (**G**) and numerator (**A**) relationship matrices for genotyped chickens

Quality control level	**G** – **A** coefficient measure	Number of animal pairs	Minimum	Maximum	Mean	Standard deviation
Strong[2]	Diagonals	4667	−0.18	0.84	0.000	0.048
	Off-diagonals	10 888 111	−0.57	1.02	0.000	0.037
	Parent-progeny pairs	5259	−0.16	0.17	−0.011	0.034
	Full-sib pairs	9126	−0.19	0.18	−0.017	0.050
	Half-sib pairs	59 870	−0.18	0.16	−0.015	0.040

Force A to be similar to G

- Christensen (2012) suggests fitting **A** to **G** instead of the opposite
  - **A** depends on pedigree completion
  - Good for chicken, bad for the rest
  - Ancestral relationships that can be seen in **G** go undetected in **A**
- Christensen analitically integrates out p_i (=allele frequencies) in a model that
  - uses $p = 0.5$ as reference in ALL loci and builds $\boldsymbol{G}_{05}$
  - uses a relationship matrix $\mathbf{A}^{\gamma}$ with <u>related</u> founders
  - The parameter γ is the relationship across founders such that we see "current" genomic relationships

Relationship across founders

Classically we assume

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

- Christensen changes this into:

$$\mathbf{A}^{\gamma} = \begin{pmatrix} 1+\frac{\gamma}{2} & \gamma & \gamma & \gamma \\ \gamma & 1+\frac{\gamma}{2} & \gamma & \gamma \\ \gamma & \gamma & 1+\frac{\gamma}{2} & \gamma \\ \gamma & \gamma & \gamma & 1+\frac{\gamma}{2} \end{pmatrix}$$

we have seen this before. The γ can also be introduced using a single metafounder

Big Data

Big Data

- Angus beef cattle:
  - 500,000 animals genotyped
  - 11M animals pedigree.
  - $\boldsymbol{A}^{-1}$ ~ 800 Mb RAM
  - $\boldsymbol{H}^{-1}$ has 350 $\times 10^9$elements ~ 2800 Gb !
- Imagine that we have to deal only with 500,000 animals genotyped
- $\boldsymbol{G}^{-1}$ is a $500{,}000 \times 500{,}000$ matrix
- $\boldsymbol{A}_{22}^{-1}$ is a $500{,}000 \times 500{,}000$ matrix
- SNP-BLUP $\boldsymbol{Z}\boldsymbol{Z}'^{-1}$ is a $50{,}000 \times 50{,}000$

Big Data

- $\boldsymbol{A}_{22}^{-1}$ can be computed efficiently using sparse matrices

$$\boldsymbol{A}_{22}^{-1} = \boldsymbol{A}^{22} - \boldsymbol{A}^{21}(\boldsymbol{A}^{11})^{-1}\boldsymbol{A}^{12}$$

See details in Yutaka et al.

Technical note: Avoiding the direct inversion of the numerator relationship matrix for genotyped animals in single-step genomic best linear unbiased prediction solved with the preconditioned conjugate gradient[1]

Y. Masuda,*[2] I. Misztal,* A. Legarra,† S. Tsuruta,* D. A. L. Lourenco,* B. O. Fragomeni,* and I. Aguilar‡

Big Data

- If number of animals > number of SNPs
- $\boldsymbol{G}$ has at most rank "number of SNPs »
  - Indirect representations of **G**
  - APY
  - Sherman-Woodbury inversions
    - $G = \frac{1}{a} I + ZZ'$
    - $G^{-1} = aI - \left(aZ \frac{1}{a} \left(\frac{1}{a} I + Z'Z \right)^{-1} Z'a \right)$

Z'Z is smaller than ZZ'

Reducing computations by ssGTBLUP

Assume: $\mathbf{G} = \mathbf{G_0} + \mathbf{C}$

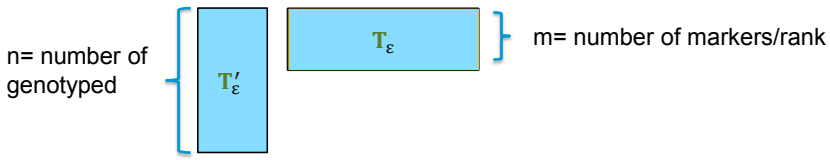
where $\mathbf{G}_0 = \mathbf{ZZ}'$ and $\mathbf{G}_\varepsilon = \mathbf{G}_0 + \varepsilon\mathbf{I}$ ➔ $\mathbf{G}_\varepsilon^{-1} = \frac{1}{\varepsilon}\mathbf{I} - \mathbf{T}_\varepsilon'\mathbf{T}_\varepsilon$

Woodbury matrix identity

where $\mathbf{T}_\varepsilon = \frac{1}{\varepsilon}\mathbf{L}_\varepsilon^{-1}\mathbf{Z}'$ and $\mathbf{L}_\varepsilon\mathbf{L}_\varepsilon' = \frac{1}{\varepsilon}\mathbf{Z}'\mathbf{Z} + \mathbf{I}$

$\mathbf{T}_\varepsilon$ has size n x m

➔ Number of computations is 2nm instead of n^2



Size of $\mathbf{T}_\varepsilon$ matrix is the same as the original marker matrix.

ssGTBLUP gives the same solutions as ssGBLUP with $\mathbf{G}_\varepsilon^{-1}$ (e.g., Koivula et al. WCGALP 2018)

Stranden, WCGALP, 2018

APY

- Misztal showed that G matrix is redundant due to limited population size
- Some chromosomal segments are copies of others
- Then G is not full rank and has a small number of eigenvalues >0

UNIVERSITY OF GEORGIA

How large-scale genomic evaluations are possible

Daniela Lourenco

05-24-2018

Algorithm for Proven and Young (APY)

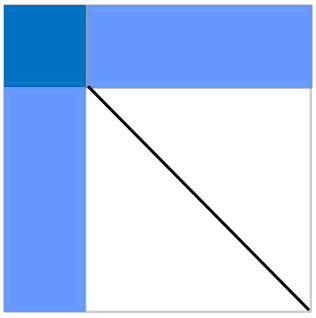
$$\mathbf{G}_{\mathrm{APY}}^{-1} = \begin{bmatrix} \mathbf{G}_{cc}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn} \\ \mathrm{I} \end{bmatrix} \mathbf{M}_{nn}^{-1} [-\mathbf{G}_{nc}\mathbf{G}_{cc}^{-1} \quad \mathrm{I}]$$

$$\mathrm{M}_{nn} = \mathrm{diag}\{\mathrm{g}_{ii} - \mathrm{g}_{ic}\mathbf{G}_{cc}^{-1}\mathrm{g}_{ci}\}$$

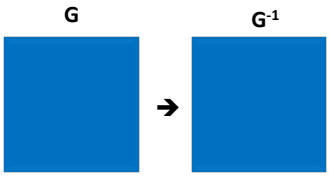
USA_287506
USA_287507
USA_287508
USA_287509
USA_287510
USA_287511
USA_287513
USA_287514
USA_287515
USA_287516
USA_287518
USA_287519
USA_287520
USA_287521
USA_287522
USA_287523
USA_287524
USA_287525
USA_287526
USA_287527
USA_287528
USA_287529
USA_287530

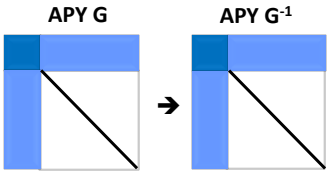
core

non-core



Algorithm for Proven and Young (APY)





- APY $\mathbf{G}^{-1}$ sparse
- Efficient computation
- Why does it work?

APY and dimension of G

genotyped animals > # SNP

$G = \alpha G + (1-\alpha)A_{22}$ VanRaden (2008)

G has a limited dimensionality
- independent blocks
- Dependent blocks

↓

Dimension of G = min (#animals, # independent SNP, *Me*)

How many core animals in APY?

largest eigenvalues of G explaining 98% ~ 99% variance

14k ~ 19k

11k ~ 16k

11k ~ 14k

4k ~ 6k

4k ~ 6k

APY and why some people don't like it

- Works well but open questions: how to choose core? Is it an approximation? etc

- Dairy cattle breeders use "Indirect Predictions" a lot
  - Estimate SNP effects every 3-4 months
  - For young animals: GEBV= sum of SNP effects , every week
- Dairy cattle breeders may prefer to work with marker effects because they use marker effects weekly: SS-SNPBLUP

SS-SNPBLUP=SSGBLUP with marker effects

Legarra & Ducrocq 2012 described a SSGBLUP model on marker effects $\boldsymbol{a}$ and BV $\boldsymbol{u}$.

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}_1'\mathbf{W}_1 & \mathbf{X}_2'\mathbf{W}_2\mathbf{Z} & \mathbf{0} \\ \mathbf{W}_1'\mathbf{X}_1 & \mathbf{W}_1'\mathbf{W}_1 + \alpha_u\mathbf{A}^{11} & \alpha_u\mathbf{A}^{12}\mathbf{Z} & \mathbf{0} \\ \mathbf{Z}'\mathbf{W}_2'\mathbf{X}_2 & \alpha_u\mathbf{Z}'\mathbf{A}^{12} & \mathbf{Z}'\mathbf{W}_2'\mathbf{W}_2\mathbf{Z} + \alpha_u\mathbf{Z}'\mathbf{A}^{22}\mathbf{Z} + \mathbf{D}^{-1}\sigma_e^2 & \alpha_u\mathbf{Z}' \\ \mathbf{0} & \mathbf{0} & \alpha_u\mathbf{Z} & \alpha_u\mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}_1 \\ \hat{\mathbf{g}} \\ -\hat{\boldsymbol{\varphi}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}_1'\mathbf{y}_1 \\ \mathbf{Z}'\mathbf{W}_2'\mathbf{y}_2 \\ \mathbf{0} \end{bmatrix}.$$

Non genotyped animals

Marker effects

Matrices **ZW** in this model get very complicated for complex models because they involve formidable products

SS-SNPBLUP=SSGBLUP with marker effects

- The model was rediscovered by Fernando et al. 2016 with the name "super hybrid model"

Non genotyped animals

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'_m\mathbf{Z}_m & \mathbf{X}'_g\mathbf{W}_v \\ \mathbf{Z}'_m\mathbf{X}_m & \mathbf{Z}'_m\mathbf{Z}_m + \mathbf{A}^{mm}\frac{\sigma_e^2}{\sigma_a^2} & \mathbf{A}^{mg}\mathbf{R}\frac{\sigma_e^2}{\sigma_a^2} \\ \mathbf{W}'_v\mathbf{X}_g & \mathbf{R}'\mathbf{A}^{gm}\frac{\sigma_e^2}{\sigma_a^2} & \mathbf{W}'_v\mathbf{W}_v + \mathbf{I}\frac{\sigma_e^2}{\sigma_\alpha^2} + \mathbf{Q}_v\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}_m \\ \hat{\mathbf{v}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'_m\mathbf{y}_m \\ \mathbf{W}'_v\mathbf{y}_g \end{bmatrix},$$

Marker effects

for same variance components,
SSGBLUP=SS-SNP-BLUP

Matrices **W** in this model get very complicated for complex models because they involve formidable products

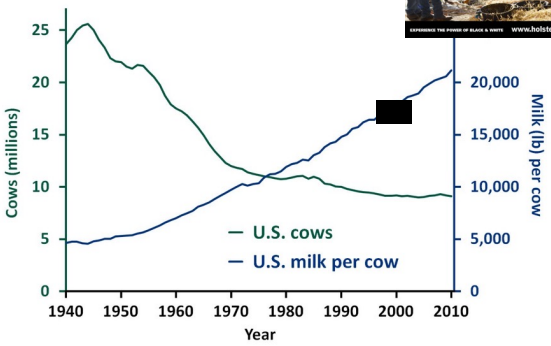
Big Data

- Large number of animals is a problem only for 1% of the users
- It is possible to fit enormous data sets with millions of genotyped animals
- The exact strategy may depend on the problem. Generality, elegance or efficiency?
- Maybe in 10 years all animals are genotyped, old data is forgotten ☺

Details in SSGBLUP

- Storage
- Inbreeding
- G is not invertible (« blending »)
- G might not explain all genetic variance (« blending »)
- Compatibility of G and A22
  - Assumption p($\mathbf{u}_2$)=N(**0**,**G**)
    - If there is selection, mean is not **0** (« tuning » solves it: see Vitezica later)
  - Same genetic variance in genotyped and ungenotyped animals
- Unknown parent groups
  - Need to modify **H** to include them (Misztal et al., 2013)
  - Metafounders
- Crosses
- Computation
  - APY
  - Sherman-Woodbury
  - « hybrid » model

Metafounders & Unknown Parent Groups



U.S. dairy population and milk yield

In the 70's there was a massive export of US Holstein to "European Friesian"

Unknown Parent Groups

- US Holstein had no data in European countries
- But treating them as equal to European cows was unfair
- European Genetic evaluations included effect of "origin"
- This effect mutated to Unknown Parent Groups

Unknown Parent Groups

- (Thompson 1979, Quaas 1988):
- Regression on % of origin computed from pedigree
  - E.g. one cow is 15% US, 80% European, 5% New Zealand
- Final EBV = portions of UPG + random part
  - $\hat{\boldsymbol{u}} = \boldsymbol{Q}\hat{\boldsymbol{g}} + \widehat{\boldsymbol{u}^*}$
  - $\boldsymbol{Q}$ contains fractions
  - $\boldsymbol{g}$ is fixed
  - $\boldsymbol{g}$ has no quantitative genetics interpretation or ”a priori” distribution
- Use of unknown parent groups is essential to get unbiased estimates across origins (UY vs US) and years (2000 vs. 2008)

Unknown Parent Groups

Unknown Parent Groups are used extensively to model:

- Missing parentship, as in sheep (father is often unknown). Genetic Groups are often defined by year of birth to model genetic progress.
- Importations, or introduction of foreign material (as in pig companies). Genetic Groups are often defined by country of origin.
- Crosses (e.g. Angus x Gelbvieh). Genetic Groups are often defined by breed.

Unknown Parent Groups in Single Step GBLUP

- Things get complicated

$$p(\boldsymbol{u}) = N(\boldsymbol{Qg}, \boldsymbol{A}\sigma_u^2)$$

$$p(\boldsymbol{u}_2) = N(\boldsymbol{0}, \boldsymbol{G}\sigma_u^2)$$

- Contradictions
- Reports of problems in SSGBLUP with complex UPG structure

Unknown-parent groups in single-step genomic evaluation

I. Misztal[1], Z.G. Vitezica[2], A. Legarra[3], I. Aguilar[4] & A.A. Swan[5]

J. Dairy Sci. 105
https://doi.org/10.3168/jds.2021-20293

Invited review: **Unknown-parent groups and metafounders in single-step genomic BLUP**

Yutaka Masuda,[1*]† Paul M. VanRaden,[2] Shogo Tsuruta,[1] Daniela A. L. Lourenco,[1] and Ignacy Misztal[1]

Unknown Parent Groups in Single Step GBLUP

J. Dairy Sci. 105
https://doi.org/10.3168/jds.2021-20293

Invited review: **Unknown-parent groups and metafounders in single-step genomic BLUP**

Yutaka Masuda,[1*]† Paul M. VanRaden,[2] Shogo Tsuruta,[1] Daniela A. L. Lourenco,[1] and Ignacy Misztal[1]

- Still open problem
- Current options
- Simplify your model !!!
- *Truncate pedigree and data*
- *Approximate UPGs* $\boldsymbol{H}^* = \boldsymbol{A}^* + \begin{pmatrix} 0 & 0 \\ 0 & \boldsymbol{G}^{-1} - \boldsymbol{A}_{22}^{-1} \end{pmatrix}$
  - $\boldsymbol{A}^*$ includes UPG using existing theory
  - $\boldsymbol{A}_{22}$ is constructed "as if" UPG don't exist, which is an approximation
  - Default in blupf90
- *Fitting UPG as covariates*
  - $\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Qg} + \boldsymbol{Wu} + \boldsymbol{e}$ with $\boldsymbol{H}^{-1} = \boldsymbol{A}^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & \boldsymbol{G}^{-1} - \boldsymbol{A}_{22}^{-1} \end{pmatrix}$
  - Final EBVs $\boldsymbol{Q}\hat{\boldsymbol{g}} + \hat{\boldsymbol{u}}$
- *Fitting "exact UPGs"*
  - Equivalent to *Fitting UPG as covariates*
  - Still not quite perfect
- Fitting « UPGs in A not in G »
- See

- The fancyest solution is « metafounders »

- The **G** matrix
  - Is exact, independently of pedigree depth
- Breeds/UPGs were considered unrelated, but they ARE related if we look at markers
- We may need to adjust the UPG theory to match **A** to **G** instead of viceversa

Missing pedigree

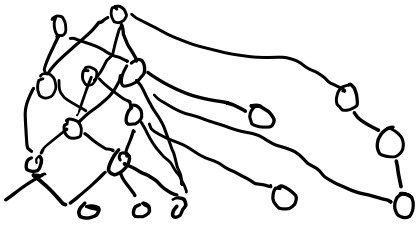
- We needed **A** to be complete
- To my knowledge, the only complete livestock pedigrees are in rabbit
- Incompleteness depend on species

- Sometimes you know the pedigree but not the associated record, so pedigree is useless

Missing pedigree

- Dairy cattle

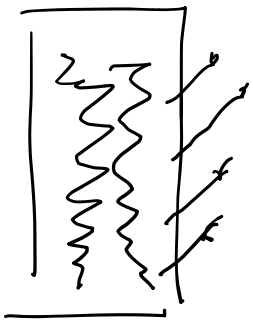
Complete for bulls and elite cows, incomplete for “cheap cows”



Missing pedigree

- Dairy sheep

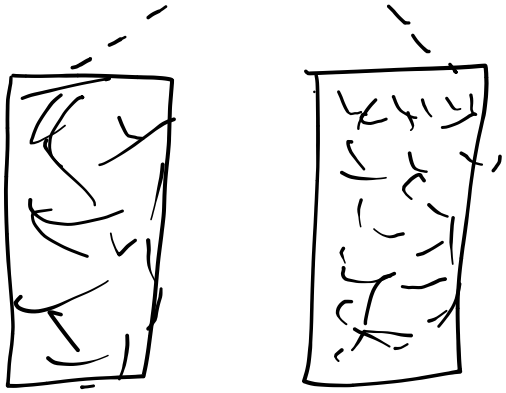
30-80% complete
Females have (often) dam known and (sometimes) sire known
Males have both parent known



Missing pedigree

If we could go back to 1700 …

- Two breeds



Covariates to fit A to G and to fit UPGs

- Hsu et al. (2017) proposed to fit a J-covariate to fit the difference between pedigree and genetic bases
- It is the same as the Vitezica et al (2010) method but fixed instead of random
- In theory the method can be extended to several populations (breeds)
  - Covariates to account for different genetc bases at G-A across breeds
  - Covariates for UPGs
  - It gets quite complicated

Strandén *et al. Genetics Selection Evolution* *(2022) 54:38*

Single-step genomic BLUP with genetic groups and automatic adjustment for allele coding

Ismo Strandén[1*], Gert P. Aamand[2] and Esa A. Mäntysaari[1]

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{0} & \mathbf{0} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{0} & \mathbf{Q}_c{}'\mathbf{G}^{-1}\mathbf{Q}_c\sigma_a^{-2} & -\mathbf{Q}_c{}'\mathbf{G}^{-1}\mathbf{Q}_2\sigma_a^{-2} & -\mathbf{Q}_c'\mathbf{F}'\sigma_a^{-2} \\ \mathbf{0} & -\mathbf{Q}_2{}'\mathbf{G}^{-1}\mathbf{Q}_c\sigma_a^{-2} & \mathbf{Q}'\mathbf{H}^{-1}\mathbf{Q}\sigma_a^{-2}+\mathbf{S}^{-1} & -\mathbf{Q}'\mathbf{H}^{-1}\sigma_a^{-2} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{X} & -\mathbf{F}\mathbf{Q}_c\sigma_a^{-2} & -\mathbf{H}^{-1}\mathbf{Q}\sigma_a^{-2} & \mathbf{W}'\mathbf{R}^{-1}\mathbf{W}+\mathbf{H}^{-1}\sigma_a^{-2} \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{c}} \\ \widehat{\mathbf{g}} \\ \widehat{\mathbf{a}}_d \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix},$$

Eq for difference of bases in G and A

Eq for UPGs

- I find this to be
  - Complicated
  - How do you define groups
  - How do you ensure that all these groups are estimable

Metafounders

- Define clusters of missing parents and call them “metafounders”
- Metafounders have relationships γ_{ij}: $\boldsymbol{\Gamma}$
- The relationships Γ_{ij} are the average relationships across missing parents of cluster i and j
- “relative to a population of maximum heterozygosity” (= identical to the “making $\boldsymbol{A}$ resemble $\boldsymbol{G}_{\mathbf{05}}$” of Christensen 2012)

To condensate:

- Things are easier if we define pseudo-individuals (metafounders) that represent "pools" of founder individuals
- These pools have self-relationships and across-relationships contained in a matrix $\mathbf{\Gamma}$.
- For instance $\mathbf{\Gamma}\begin{pmatrix} Holstein \\ Jersey \end{pmatrix} = \begin{pmatrix} 0.55 & 0.48 \\ 0.48 & 0.77 \end{pmatrix}$
- Holstein is more variable than, and related to, Jersey
- Build **A** from $\mathbf{\Gamma}$ following tabular rules

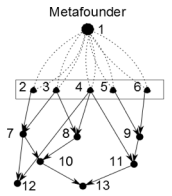
Metafounder relationships

RELATIONSHIPS

Across founders within the population

Across founders across the populations

A SINGLE METAFOUNDER

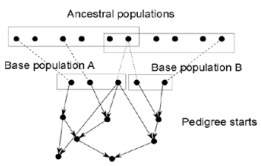


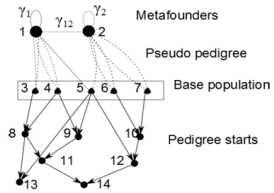
Pedigree
1 0 0
2 1 1
3 1 1
4 1 1
5 1 1
6 1 1
7 2 3
8 3 4
9 5 6
10 7 8
11 4 6
12 7 4
13 10 11

$$\mathbf{A}^{\gamma} = \begin{pmatrix} \gamma & \gamma & \gamma & \gamma \\ \gamma & 1+\frac{\gamma}{2} & \gamma & \gamma \\ \gamma & \gamma & 1+\frac{\gamma}{2} & \gamma \\ \gamma & \gamma & \gamma & 1+\frac{\gamma}{2} \end{pmatrix}$$

It has self-relationship $A_{11} = \gamma$ so $F = \gamma - 1$.
If $\gamma = 0$ then we have regular relationships.
All **A** and $\mathbf{A}^{-1}$ methods work.

TWO OR MORE METAFOUNDERS





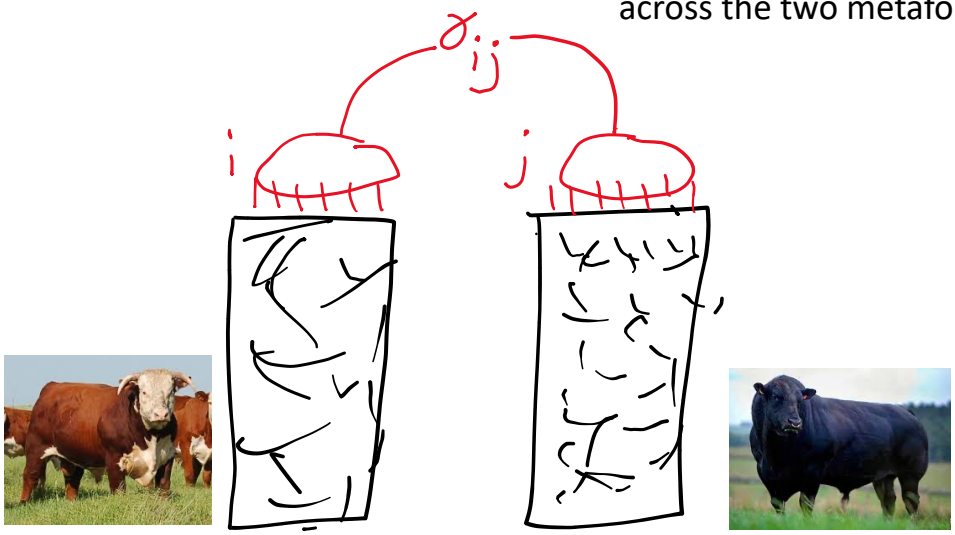
$$\mathbf{\Gamma} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \\ & \gamma_{22} & \\ & & \ldots \end{pmatrix}$$

Pedigree
1 0 0
2 0 0
3 1 1
4 1 1
5 1 2
6 2 2
7 2 2
8 3 4
...

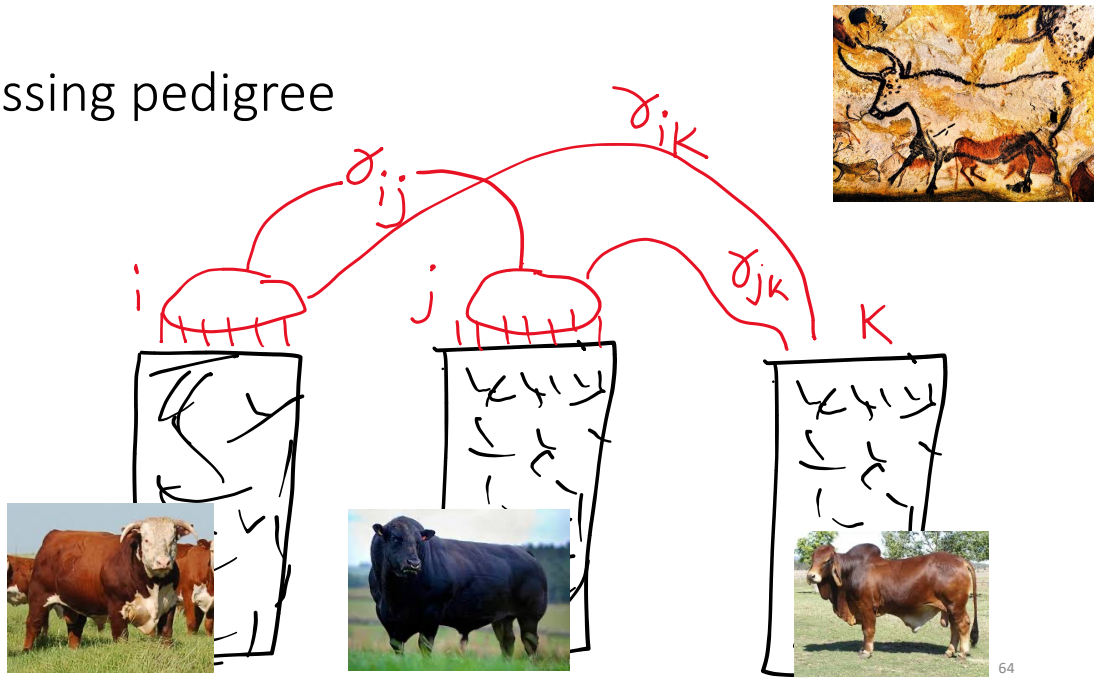
Algorithms change but they are still easy.

Missing pedigree

- We define a relationship γ_{ij} across the two metafounders

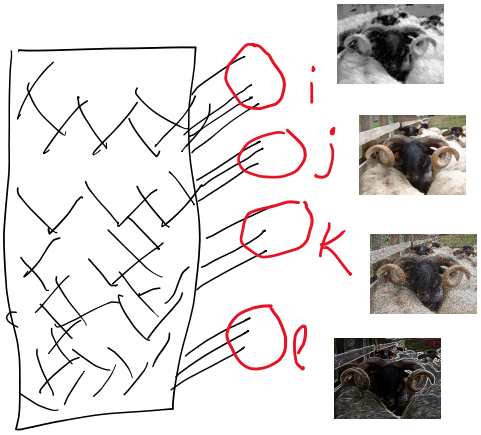


Missing pedigree



Missing pedigree

- One breed



$$\boldsymbol{\Gamma} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} & \gamma_{24} \\ \gamma_{31} & \gamma_{32} & \gamma_{33} & \gamma_{34} \\ \gamma_{41} & \gamma_{42} & \gamma_{43} & \gamma_{44} \end{pmatrix}$$

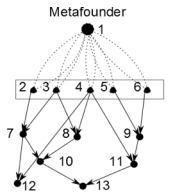
Metafounder relationships

RELATIONSHIPS

Across founders within the population

Across founders across the populations

A SINGLE METAFOUNDER

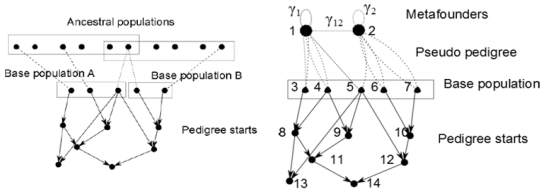


Pedigree
1 0 0
2 1 1
3 1 1
4 1 1
5 1 1
6 1 1
7 2 3
8 3 4
9 5 6
10 7 8
11 4 6
12 7 4
13 10 11

$$\mathbf{A}^{\gamma} = \begin{pmatrix} \gamma & \gamma & \gamma & \gamma \\ \gamma & 1+\frac{\gamma}{2} & \gamma & \gamma \\ \gamma & \gamma & 1+\frac{\gamma}{2} & \gamma \\ \gamma & \gamma & \gamma & 1+\frac{\gamma}{2} \end{pmatrix}$$

It has self-relationship $A_{11} = \gamma$ so $F = \gamma - 1$.
If $\gamma = 0$ then we have regular relationships.
All **A** and **A**$^{-1}$ methods work.

TWO OR MORE METAFOUNDERS



$$\boldsymbol{\Gamma} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \\ & \gamma_{22} & \\ & & \ldots \end{pmatrix}$$

Pedigree
1 0 0
2 0 0
3 1 1
4 1 1
5 1 2
6 2 2
7 2 2
8 3 4
...

Algorithms change but they are still easy.

Inbreeding with metafounders

- Assume
  - $\gamma_{BB2012-BB2012} = 0.2$
  - $\gamma_{BB2014-BB2014} = 0.1$
  - $\gamma_{BB2012-BB2014} = 0.04$
- Then
  - "Lucy" BB animal born in 2012 with unknown parents has F=0.1
  - "Sean" BB animal born in 2014 with unknown parents has F=0.05
  - Relationship of Sean and Lucy is 0.04
- "Paul" offspring of Sean and Lucy has inbreeding 0.02
- This shows that we compensate for missing pedigrees

Metafounders

- Metafounders have relationships γ_{ij}: $\boldsymbol{\Gamma}$
- $\gamma_{ij} = 8Cov(p^i, p^j) = 8\sigma_{p_b,p_{b'}}$ with *p* at each base opulation
- Related to Fst differentiation indices and to genetic and evolutionary distances

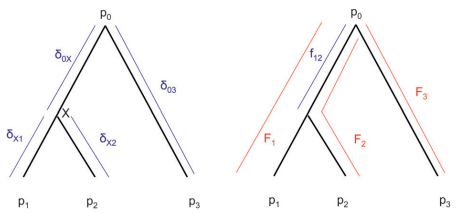


FIGURE 1.—Example of tree-like evolution: construction of the kinship matrix.

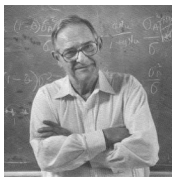
The kinship matrix: Due to drift and coancestries, frequencies p_i's are correlated, so that

$$\mathrm{Cov}(p_i, p_j) = f_{ij} p_0 (1 - p_0) \quad (1)$$

$$\mathrm{Var}(p_i) = f_{ii} p_0 (1 - p_0), \quad (2)$$

Detecting Selection in Population Trees: The Lewontin and Krakauer Test Extended

Maxime Bonhomme,* Claude Chevalet,* Bertrand Servin,* Simon Boitard,* Jihad Abdallah,*,† Sarah Blott‡ and Magali SanCristobal‡,1



- Cockerham (1969) and Robertson (1975) interpret $4\sigma_{p_b,p_{b'}}$ as the coancestry across two populations and Fariello et al. (2013) use $\sigma_{p_b,p_{b'}}$ to describe the divergence of populations.
- There are several measures of genetic distance between populations (e.g. (Laval *et al.* 2002)), and most of them contain a term related, implicitly or explicitly, to $\sigma_{p_b,p_{b'}}$.
- It is also related to Fst and Nei's distance (see extra doc)

Compatibility of G and A using metafounders

- ✓ Extension of Christensen (2012)
- ✓ Write as many metafounders as base populations
- ✓ These metafounders are related by a matrix of additive relationships $\boldsymbol{\Gamma}$
- ✓ Estimate $\boldsymbol{\Gamma}$ using markers and pedigree (and maybe data)
- ✓ Define $\mathbf{G}$ as crossproduct $\mathbf{G} = \frac{(\boldsymbol{M}-2\boldsymbol{P})(\boldsymbol{M}-2\boldsymbol{P})'}{\frac{n}{2}}$ with **P** containing 0.5
- ✓ Then combine everything into <u>one</u> H matrix for all animals

$$\mathbf{H}^{\boldsymbol{\Gamma}^{-1}} = \mathbf{A}^{\boldsymbol{\Gamma}^{-1}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{\boldsymbol{\Gamma}^{-1}} \end{bmatrix}$$

- $\mathbf{A}^{\boldsymbol{\Gamma}^{-1}}$: first invert $\boldsymbol{\Gamma}$, then use Henderson's rules
- This is the "best" compatibility of **G** and **A**

Technical note: Genomic evaluation for crossbred performance in a single-step approach with metafounders[1]

T. Xiang,*†[2] O. F. Christensen,* and A. Legarra†

Re-analyses of exact same data as previous paper:

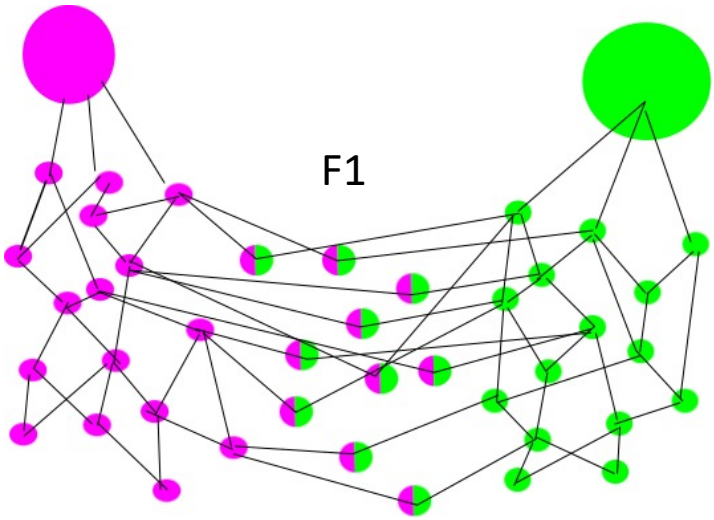
Application of single-step genomic evaluation for crossbred performance in pig[1]

T. Xiang,*†[2] B. Nielsen,‡ G. Su,* A. Legarra,† and O. F. Christensen*

Landrace

Yorkshire

F1



Some are genotyped (7700 L, 7700 Y, 5500 F1)

TNB was recorded in 293,339 LL, 180,112 YY, and 10,974 crossbred.

332,929 LL, 210,554 YY, and 10,974 crossbreds were in the pedigree

Landrace x Yorkshire = F1 (Tao Xiang)

- Single Step
- Genotypes and phenotypes in purebreds and crosses
- Old method: two SSGBLUPs separate for each origin (Xiang 2016 J Anim Sci)
- New method: metafounders
- Two populations Landrace and Yorkshire

$$\mathbf{\Gamma} = \begin{bmatrix} \hat{\gamma}_L & \hat{\gamma}_{L,Y} \\ \hat{\gamma}_{L,Y} & \hat{\gamma}_Y \end{bmatrix} = \begin{bmatrix} 0.756 & 0.259 \\ 0.259 & 0.730 \end{bmatrix}$$ estimated by GLS

Landrace x Yorkshire = F1 (Tao Xiang)

- <u>One</u> H matrix for <u>all</u> animals (Landrace, Yorkshire, or F1)

$$\mathbf{H}^{\mathbf{\Gamma}^{-1}} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{\mathbf{\Gamma}^{-1}} \end{bmatrix} + \mathbf{A}^{\mathbf{\Gamma}^{-1}},$$

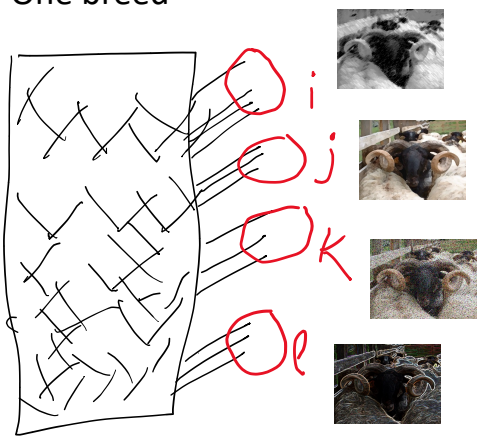
- <u>Three trait</u> model (L,Y, F1) depending on which population the trait was recorded
- The three trait model accommodates interactions GxG and GxE.

Landrace x Yorkshire = F1 (Tao Xiang)

- The results were as good as the more complex method in the previous paper
- But much easier

Missing pedigree

- One breed



$$\boldsymbol{\Gamma} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} & \gamma_{24} \\ \gamma_{31} & \gamma_{32} & \gamma_{33} & \gamma_{34} \\ \gamma_{41} & \gamma_{42} & \gamma_{43} & \gamma_{44} \end{pmatrix}$$

J. Dairy Sci. 105
https://doi.org/10.3168/jds.2021-20860

Removing data and using metafounders alleviates biases for all traits in Lacaune dairy sheep predictions

F. L. Macedo,[1,2,3] J. M. Astruc,[4] T. H. E. Meuwissen,[5] and A. Legarra[1*]

Use of UPG results in similar estimates than using MF
...but more biased evaluations (b_p <<1)

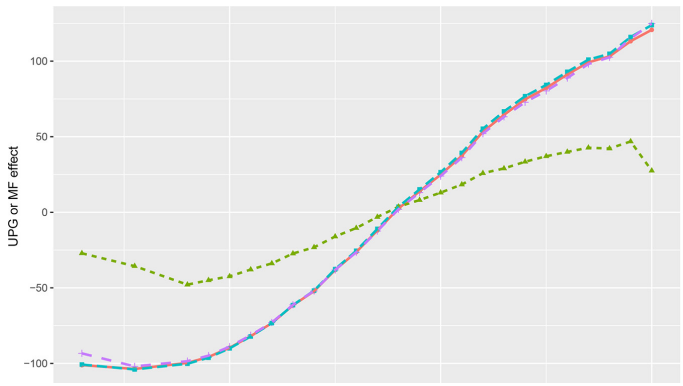


Figure 4. Estimates of unknown parent groups (UPG) and metafounders effect for milk yield

Table 5. Slope ($\hat{b}_p$) of the regression of EBV_w (whole data set) on EBV_p (

Evaluation[1]	Model[2]	MY
Official	BLUP-UPGA	0.71
Official	SSGBLUP-MF-trend	0.86
Official	SSGBLUP-MF-peeling	0.75
Official	SSGBLUP-UPGH	0.59

UPG or metafounders?

- UPG are “fixed effects”
- Metafounders are “random and related” effects

- There are equations to use UPG as fixed effects in SSGBLUP
  - but they’re more complex to implement than metafounders
  - The compatibility of G and A using “J-coefficients” needs to be obtained separately
- Metafounders
  - does compatibility & estimation in one shot
  - computationally very simple
  - estimation of $\boldsymbol{\Gamma}$ is tricky

State of the art for Metafounders

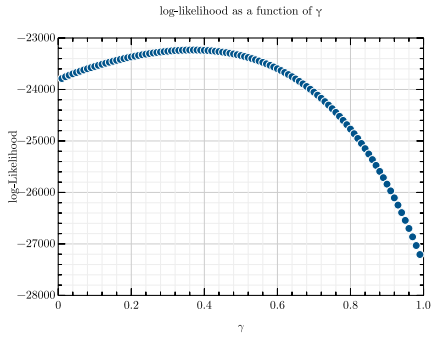
- Right now there are a few papers using MF, in particular (to my knowledge) in pure & crossbred dairy cattle and sheep
- A complex issue is how to estimate (or assume) $\gamma_{ij} = 8Cov(p^i, p^j) = 8\sigma_{p_b,p_{b'}}$ with *p* at each base opulation
- It is similar, but not equal, to estimate a variance component
- in particular if MF are distant from genotypes
- We have now:
  - A Maximum Likelihood estimate for a single γ
  - A method using increase of inbreeding for multivariate $\mathbf{\Gamma}$ within breed (this is based in Macedo et al. 2021)
- See extra docs for details

State of the art for Metafounders

- Single γ
- Maximum Likelihood
- get $\mathbf{G_{05}} = \mathbf{MM}/(k/2)$ with M={-1,0,1} and k number of markers
- get $\boldsymbol{A_{22}}$
- compute $a = \mathbf{1}'\mathbf{A_{22}^{-1}}\mathbf{1}$, $b = Tr(\mathbf{A_{22}^{-1}}\mathbf{G})$ and $c = Tr(\mathbf{A_{22}^{-1}}\mathbf{1}\mathbf{1}'\mathbf{A_{22}^{-1}}\mathbf{G})$
- the ML estimate of γ is the solution of a cubic equation

Test with Lacaune AI rams (which have very complete pedigrees)

the value of $\hat{\gamma} \approx 0.46$ makes pedigree and genomic relationships "most compatible"



State of the art for Metafounders

- Multiple γ, simple population
- We conceived a method based on average increase of relationship in the population
  - Average relationships in t generations a population increase by $2\Delta F$

$$\overline{\mathbf{A}}_{(0:t,0:t)} = \begin{pmatrix} \overline{A}_0 & \overline{A}_0 & \overline{A}_0 & \dots \\ \overline{A}_0 & \overline{A}_1 & \overline{A}_1 & \dots \\ \overline{A}_0 & \overline{A}_1 & \overline{A}_2 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & \dots \\ 0 & 2\Delta F & 2\Delta F & \dots \\ 0 & 2\Delta F & 4\Delta F & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix} = 2k\mathbf{TT}'\Delta F; \quad \mathbf{T} = \begin{pmatrix} 0 & 0 & 0 & \dots \\ 1 & 0 & 0 & \dots \\ 1 & 1 & 0 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}'$$

  - If missing parents are drawn from the general population at random, then their relationship is also described by $\overline{\mathbf{A}}$ (VanRaden 1992).
  - Metafounders describe average relationships so we have:

$$\mathbf{\Gamma} = \begin{pmatrix} \Gamma_0 & \Gamma_0 & \Gamma_0 & \dots \\ \Gamma_0 & \Gamma_0 + 2\Delta F_{\mathbf{\Gamma}} & \Gamma_0 + 2\Delta F_{\mathbf{\Gamma}} & \dots \\ \Gamma_0 & \Gamma_0 + 2\Delta F_{\mathbf{\Gamma}} & \Gamma_0 + 4\Delta F_{\mathbf{\Gamma}} & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix} = \mathbf{11}'\Gamma_0 + 2k\mathbf{TT}'\Delta F(1 - \Gamma_0/2)$$

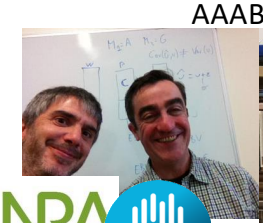
  - we got good results in Lacaune (good alignment of pedigree and genomic relationships"
- the remaining case is "breeds + crosses + missing pedigree" 🤔
- Kudinov et al., 10.3389/fgene.2022.1012205 ; Legarra et al. Gen Sel Evol (2024) 56:35

Method LR in theory and practice

Legarra, A., Reverter, A., Macedo, F

andres.legarra@inra.fr

Financed by an INRA-CSIRO joint action, program Genopyr and AAABG,trips to UGA, and ARDI grant and Smarter



AL: INRAE

AR: CSIRO (Australia)

FM: Universidad de la Republica (Uruguay), INRAE, Interbull (Sweden)

Risks of forecasting in Animal Breeding

- Spelman et al:
  - New Zealand Animal Evaluation Limited (NZAEL), […] first included genomic information in the national evaluation in 2009 […]
  - The first two crops of DNA-proven sires, used in 2008 and 2009 […] the initial GEBVs of these sires were found to be over-estimated […] as a gesture of appreciation to the early adopters of genomic evaluation, LIC credited the $5 premium that the farmers paid
- (Similar report by Sargolzaei, Chesnais…)
- We need tools to rank, understand and quantify the behavior of prediction models in an “animal breeding” context
- The need for these tools has dramatically increased with genomic selection, that takes riskier decisions

Why do we need cross-validation?

- Classic statistical methods to compare models (AIC…) only inform about global fit and parsimony of ALL data
- (In Animal Breeding) we want our models to predict THE FUTURE offspring of selected animals
- We're not interested in better finding out the effect that "lambing at 18 months" had in 1998 – we want to know the best sheep NOW

Cross-validation in a nutshell

- Split data into "training" and "validation",
- using a model and data in "training" predict "data" in "validation"
- Measure quality of prediction

Which kind of cross-validation should we use?

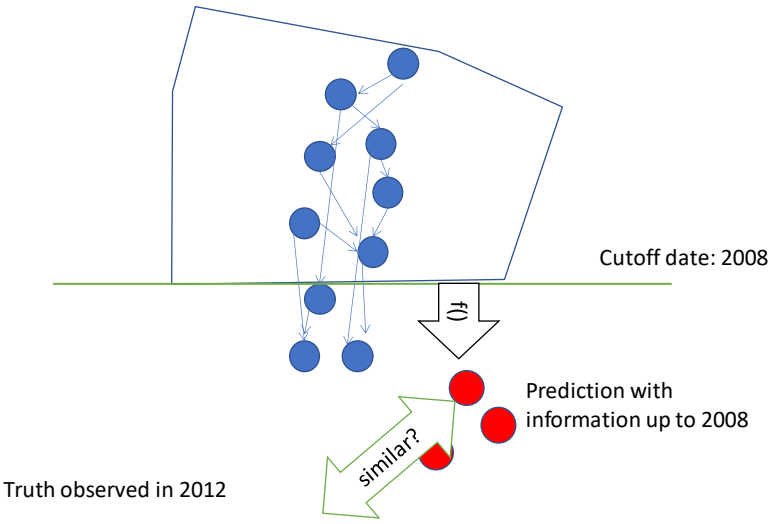
For selection:

- Is my genetic evaluation leading me to maximization of genetic progress?
- We want the method that best predicts <u>future</u> performance
- Forward cross-validation (or retrospective analysis)
  - (Interbull tests)
  - Cut data at date *t*
  - Could we have predicted at time *t* the data that was actually observed <u>after</u> *t*?

Which kind of cross-validation should we use?

- Random k-fold, leave-one-out, k-means cluster for crossvalidation:
  - Predict contemporaries (and not offspring)
  - It might be useful to predict performance of existing individuals in, say, other environments (plants)
  - Results from these crossvalidations should *not* be taken as "reliabilities" in a selection theory sense
- Random k-fold: you may predict e.g. parents from offspring
- Leave-one-out: overfit
  - (it has been proven in Stat literature)
  - there's always a close sib with information
  - and are we interested in predicting well one individual?

Example: Forward cross-validation



Metrics

- Theory of quantitative genetics suggest using Metrics from linear regression of u (TBV) on $\hat{u}$ (EBV):

- Bias: $b_0 = E(u - \hat{u})$ (it is NOT the intercept of the regression of u on $\hat{u}$)
- Slope: $b_1 = \frac{Cov(u,\hat{u})}{Var(\hat{u})}$ (slope of the regression of u on $\hat{u}$)
- Accuracy: $r = \frac{Cov(u,\hat{u})}{\sqrt{Var(u)Var(\hat{u})}}$

In fact: $MSE = b_0^2 + \sigma_u^2 \left(1 + \frac{r^2}{b_1^2} - \frac{2r^2}{b_1}\right)$

- Why are these relevant? Genetic progress !!

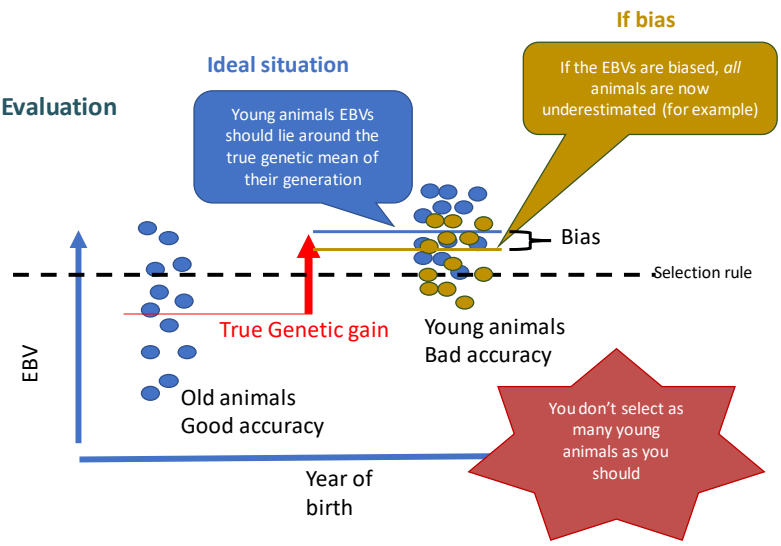
True and estimated genetic progress

- When we select animals, we *believe* our $\Delta G = \frac{1}{n}\Sigma(GEBV) = \bar{\hat{u}}$
- This only holds if bias $b_0 = 0$, regression $b_1 = 1$
- $b_0 < 0$ or $b_1 < 1$ (overdispersion) lead to overestimation of selected young animals

- So, in addition to accuracy r we should check both b_0 and b_1

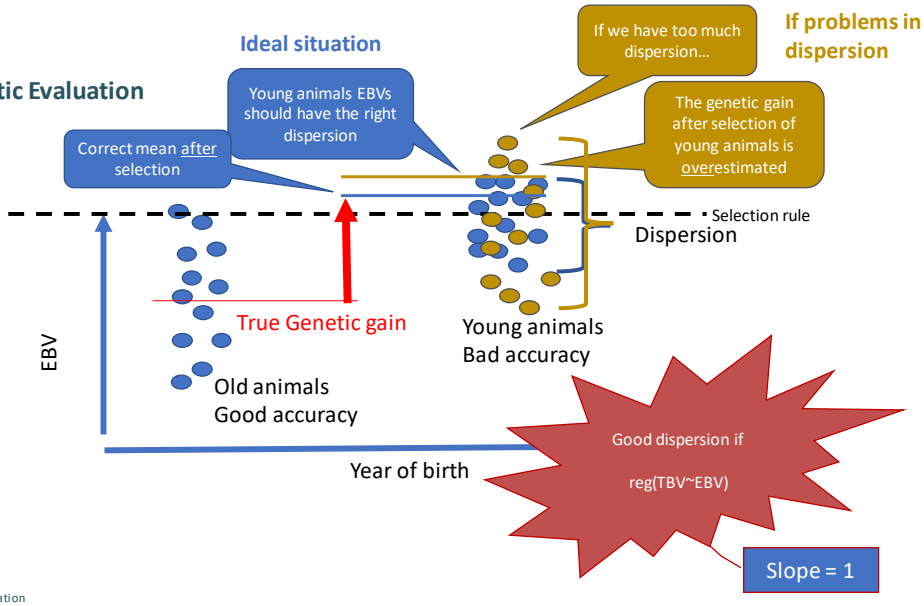
Genetic gain: b_0

Consider a Genetic Evaluation



Dispersion: b_1

Consider a Genetic Evaluation



How should we cross-evaluate?

- We can see $\hat{u}$ (EBVs), we can't see u (TBVs)
- "Predictivity": compare predictions with observations (precorrected phenotypes y^* or deregressed proofs) :
  - e.g. $r \approx cor(y^*, \hat{y})/h$ (Legarra et al. 2008)
- But this ignores the covariance structure in precorrected $y^*$and leads to paradoxes:
  - $r > 1$ (observed in chicken)
  - $r_{pedigree} > r_{genomic}$ (observed in dairy cattle for fertility)
- It also ignores that candidates to selection have reduced genetic variance
- [for this: see page 9 in Legarra-Reverter 2018]

Underestimation of accuracy using predictability due to reduction of genetic variance

- $r\left(\boldsymbol{y}^*_{new}, \widehat{\boldsymbol{u}}_p\right)$ has expectation $\frac{acc}{h}$
  - only when animals are NOT selected
- often is not the case: for instance, prospective AI rams (jeunes agneaux), their genetic variance $\sigma^2_{u^*}$ is less than the "normal" genetic variance σ^2_u
- Using $\widehat{acc} \approx \frac{\rho_{y,p}}{h}$ underestimates populational accuracy

Fictional example in dairy cattle

- Using the "dairy" example in [Bijma 2011] …
- assume observed $\rho_{y_c,p} = 0.3$,
  - this yields (biased) $\widehat{acc} \approx \frac{\rho_{y,p}}{h} = 0.55$ and
  - (correct) $\widehat{acc} \approx \frac{\rho_{y,p}}{h_\infty} = 0.67$.
  - This value can, in turn, be translated as an "unselected accuracy" of 0.82

Overestimation of accuracy using predictability due to ignoring error in estimate of fixed effects

- We use $\boldsymbol{y}^*_{new}$ as it was "exact"
- For a balanced design with n_i records per contemporary group
- $\frac{E\left(\rho_{y^*_{new},p}\right)}{h} \approx acc_p\left(1+\frac{1}{n_i}\right)$
  - a (relative) <u>over</u>estimation of the accuracy of $acc_p\left(\frac{1}{n_i}\right)$
- Dairy sheep: 25 animals / contemporary group, overestimation of accuracy by 4%
- Beef cattle: 5 animals / contemporary group, overestimation of accuracy by 20%

How should we cross-evaluate?

- Dairy cattle breeders use DYDs (average performance of daughters after correction)
  - In other species, DYDs are very little reliable (pigs!! but also sheep and goat)
  - Analysis of DYDs assumes that they are "uncorrelated" across bulls, but this is false when the number of daughters is small or the trait is low heritable
  - Use of genomic selection makes DYDs more and more biased
- "Deregressed Proofs" (Garrick et al.) suffer the same problems as "predictivity" unless large progenies
  - (also: The method of Garrick is not quite correct, see Ricard-Legarra-Danvy JAS 2013)

Legarra & Reverter (2018) proposed a new method based on comparisons of EBV from partial (old) data vs whole (old+new) data.

- Does not require “true” breeding values
- Does not require pre-corrected phenotypes
- Could be used for any kind of traits

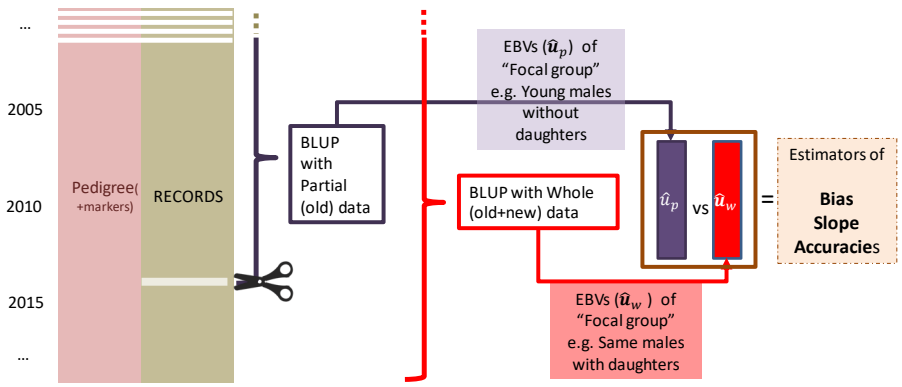
Check of bias using successive evaluations

- Legarra, A., & Reverter, A. (2018). Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genetics Selection Evolution*, *50*(1), 1-18.
- Legarra, A., & Reverter, A. (2019). Correction to: Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genetics Selection Evolution*, *51*(1), 1-2.

- We proved (analytically) that in successive genetic evaluations there are useful statistical properties of the distributions of “early” and “late” EBVs
- We use these properties to get estimators of biases and accuracies

The proposed method LR

How does LR method works?



Practicalities: defining focal groups

- The properties of the method hold for a group of animals that are contemporaries and have
  - similar information at ”partial” (e.g. only Parent Average)
  - and similar information at “whole” predictions (e.g. Parent Average + phenotype, or Parent Average + offspring, or…)
- we call this focal group
- we're interested in the group, not in each individual animal
- young born rams (bulls) can be a focal group.
- 1st-lambing females can be a focal group, and
- rams with first crop of daughters could be a focal group

Estimators of LR method: Bias and Slope

Old evaluations

New evaluations

Bias $\hat{\Delta}_p = \bar{\hat{\boldsymbol{u}}}_p - \bar{\hat{\boldsymbol{u}}}_w$

Expected value of 0 in absence of bias

Slope $\hat{b}_p = \frac{cov(\hat{\boldsymbol{u}}_p, \hat{\boldsymbol{u}}_w)}{var(\hat{\boldsymbol{u}}_p)}$

Expected value of 1 in unbiased genetic evaluations

2018

Top 50 Ranked by Net Merit $

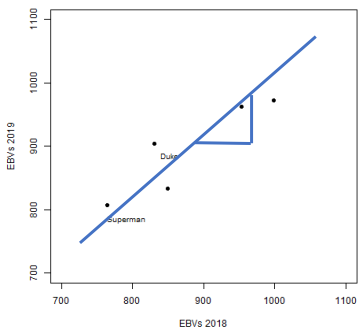
NAAB	Name	NM$	Rel.
29HO17553	JOSUPER	999 G	96
1HO11881	PRINCETON	849 G	93
250HO13267	DUKE	831 G	82
151HO681	RUBICON	953 G	95
200HO7846	SUPERMAN	764 G	93

2019

Top 50 Net Merit $

NAAB Code	Short Name	NM$	OR	Rel
29HO17553	JOSUPER *99-I	973	G	97
1HO11881	PRINCETON *99-I	833	G	95
250HO13267	DUKE	904	G	92
151HO681	RUBICON	963	G	97
200HO7846	SUPERMAN	807	G	95

$$\mu_{wp} = \frac{999 + 849 + 831 + 953 + 764}{5} - \frac{973 + 833 + 904 + 963 + 807}{5} = -16.8$$




```
EBV2018=c(999,849,831,953,764)
EBV2019=c(973,833,904,963,807)
delta_wp=mean(EBV2018)-mean(EBV2019)# -16.8
aa=lm(EBV2019~EBV2018)
b_wp=aa$coefficients[2] # 0.71
```

$\hat{\Delta}_{wp}$

$\hat{b}_p$

Estimators of LR method: Accuracies

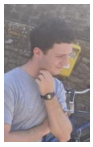
Relative estimators

Ratio of accuracies

$\hat{\rho}_{w,p} = \frac{Cov(\hat{u}_p, \hat{u}_w)}{\sqrt{Var(\hat{u}_p)Var(\hat{u}_w)}}$ with expected value $\frac{acc_p}{acc_w}$ (values close to 1 indicate that "partial evaluation" was "as accurate" as later evaluation, but both evaluations could be "little accurate")

Relative increase in accuracy

$\frac{1}{\hat{\rho}_{w,p}} - 1$ with expected value $\frac{acc_w - acc_p}{acc_p}$ (if $\frac{acc_w - acc_p}{acc_p} = 0.5$ then genetic progress increases by 50%)



Ratio of reliabilities

$\hat{\rho}^2_{p,w} = \frac{Cov(\hat{u}_p, \hat{u}_w)}{Var(\hat{u}_w)}$ with expected value $\frac{acc_p^2}{acc_w^2}$ (ratio of reliabilities)

Estimators of LR method: Accuracies

Direct estimators

Selected reliability

$$\widehat{acc}^2{}_p = \frac{Cov(\hat{\boldsymbol{u}}_p, \hat{\boldsymbol{u}}_w)}{\sigma^2_{u^*}}$$

The denominator $\sigma^2_{u^*}$ is the variance of animals in the focal group (and not the variance of the base generation).

When animals are pre-selected: for instance, prospective AI rams, their genetic variance $\sigma^2_{u^*}$ is less than the "normal" genetic variance σ^2_u

- In MTR, $\sigma^2_u \approx 500$ but $\sigma^2_{u*} \approx 350$ for young rams (milk yield)
- This equation $\widehat{acc}^2{}_p$ gives the "selected" reliability of Bijma (J. Anim. Breed. Genet. (2012) 1–14) and Dekkers (Anim Sci 1992)
- This reliability says the "ability" to rank *within* those animals (more difficult when they're selected)
- $\sigma^2_{u^*}$ can be estimated using e.g. Gibbs sampling (proven bulls is $\sigma^2_{u^*} \approx var(EBV)$)

But we can't use this accuracy for the whole population, and we can't compare it with results in less selected animals, say, beef cattle

Estimators of LR method: Accuracies

Direct estimators

- Solution: correct using ratios of unselected and observed genetic variances of these animals :

Unselected reliability

$$\widehat{rel}_p = 1 - \frac{\sigma^2_{u^*}}{\sigma^2_u}\left(1 - \widehat{acc}^2_p\right)$$

- This matches what you should get from the inverse of the MME (Model-based reliabilities)
- The mathematical explanation of all this is quite boring but is detailed in the Appendix of Macedo et al. 2020 J Dairy Sci
- The computation of $\sigma^2_{u^*}$ etc etc can be found in Macedo et al. 2020 GSE (Gibbs sampler, no problem for < 10 M animals).

Examples of estimation of accuracies (MTR)

Model	Selected reliability $\widehat{acc}_p^2$	Unselected reliability $\widehat{rel}_p$	Ratio of reliabilities $\hat{\rho}_{pw}^2$
BLUP-MF	0.22	0.53	0.32
BLUP-UPG	0.24	0.54	0.31
SSGBLUP-MF	0.32	0.59	0.45

All of them agree in saying SSGBLUP >> BLUP

- The "unselected reliability" is in the scale of Reliability
- The "ratio of reliabilities" is harder to interpret

Macedo et al. Genet Sel Evol (2020) 52:47
https://doi.org/10.1186/s12711-020-00567-1

RESEARCH ARTICLE Open A

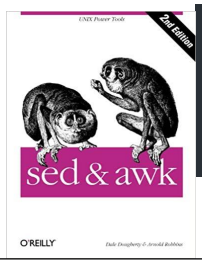
Bias and accuracy of dairy sheep evaluations using BLUP and SSGBLUP with metafounders and unknown parent groups

Practicalities: defining focal groups

- In dairy sheep we take the data file and we work looking forward
- Take all rams born in 2014 that were used in AI in the breed MTR
- Few years later (say 2017) we find out which of these rams have daughters with milk yield
- This defines a focal group for "partial"=2014 and "whole"=2017
- We can do the same for 2014 vs. 2018, 2019, etc
- lots of work of data exploring but we have

Practicalities: defining “whole” and “partial”

- You can do many “partials” and many “wholes”
- for instance you can do “partial” at 2010, 2011,…
- and compare each of them vs. “whole” at 2014, 2015…
- it is important to do several comparisons !!
- this requires automatic handling of files and data editing, fortunately we have



for instance: work in MTR

Evaluations with data until 2005, until 2006 and so on until 2017.

We compare

- **EBVs at birth** (EBV_P) of a set of Artificial Insemination males (2005 – 2014)
- **EBVs of the same males in later evaluations** (after having progeny) (EBV_W) (until 2017).

For example for males born in 2005, 11 pairs of evaluations were analysed,

- 2005 vs 2007; 2005 vs 2008; … and 2005 vs 2017

The same for males born in 2006, 2007 … 2015

- 2006 vs 2008; … and 2015 vs 2017

Total of 65 comparisons that we “average” using a linear model to account for unbalance (details in the paper)

Practicalities: defining “whole” and “partial”

- Delete records (**y**) after cut-off date
- ideally, keep pedigree and markers only up to the cut-off date
  - for instance if “partial”= March 2014 we should keep records, pedigree and markers up to March 2014 (because pedigree and markers were used to predict the young rams)
- in practice:
  - we delete “records” (milk yield, etc etc) based on the year
  - for practicality, we keep ALL pedigree and ALL markers
  - we don’t think this should give problems because pedigree does not contribute information to ssGBLUP, and genotypes very very little

Practicalities: referring to same genetic base

- In genetic evaluations with Unknown Parent Groups, the EBVs are not estimable functions
- So you need to refer all EBVs to a common genetic base in order to infer “bias” or not.
- Typically the genetic base is something like “average EBV of all females born in 2010” or something like that.

Practicalities: genomic *vs* pedigree

- Wait, how do know if I want to spend money in genotypes?
- In other words, how do I compare a "genomic" and a "non genomic" evaluation?
- Easy:
  - keep the same records (**y**) and pedigree (**A**) for two analyses:
  - "non genomic" = "partial"
  - "genomic" = "whole" (markers = more "data")
- Alternatively: run "whole" and "partial" with "genomic" and "pedigree" (4 evaluations in total)

What do we do with several models?

- If the model is correct it should be "internally" coherent (no biases, $b_0 = 0$ and $b_1 = 1$)
  - Don't use models that are obviously "internally" biased
- Can I compare two models?
  - "partial" with one model and "whole" with another model
- We did for "partial" = "old data without markers (BLUP)" vs. "whole" = "new data with markers (ssGBLUP)"
  - Because ssGBLUP requires changing the model! (a bit)
- We expect to see no biases

Some results

Fernando Macedo PhD

- Dairy sheep improvement is a French specialty !
- This is <u>not</u> meat or wool sheep
- Very well structured "mini – dairy cattle style" breeding program
- AI, performance recording, etc etc
  - <2015: progeny-testing
  - >2015: genomic selection

The breeds

Manech Tête Rousse

Female population size:	274,000
Females in the breeding flocks:	80,260 (29%)
Tested rams per year:	150
Rams at AI Center:	600
Individuals in the pedigree	540,999
Number of records Milk Yiedl:	1,842,295
Missing pedigree:	≈25%

Traits selected: Milk Yield and contents

Lacaune

Female population size:	890,000
Females in the breeding flocks:	174,472 (19%)
Tested rams per year:	440
Rams at AI Center:	1400
Individuals in the pedigree	1,868,975
Number of records Milk Yield:	5,696,348
Missing pedigree:	≈9%

Traits: Milk Yield and contents, SCS, Udder traits

Genotyping in French dairy sheep

- Every year, new lambs enter the AI center
- They have undergone two steps of selection:
  - 1st selection: based on "parent average" (½ mom + ½ dad) => retain so many (*n*) "best" lambs (say 1000) among all male lambs in the breed (say 50,000)
  - Genotype the *n* "best" lambs with 15K chip; impute
  - do genomic predictions for the *n* newly genotyped "best" lambs
  - 2nd selection: based on genomic prediction, retain the *n/5* "best best" lambs (say 200)
  - These "best best" *n/5* lambs enter the AI center

Main results Manech Tête Rousse

Some models (UPGH…) are really biased

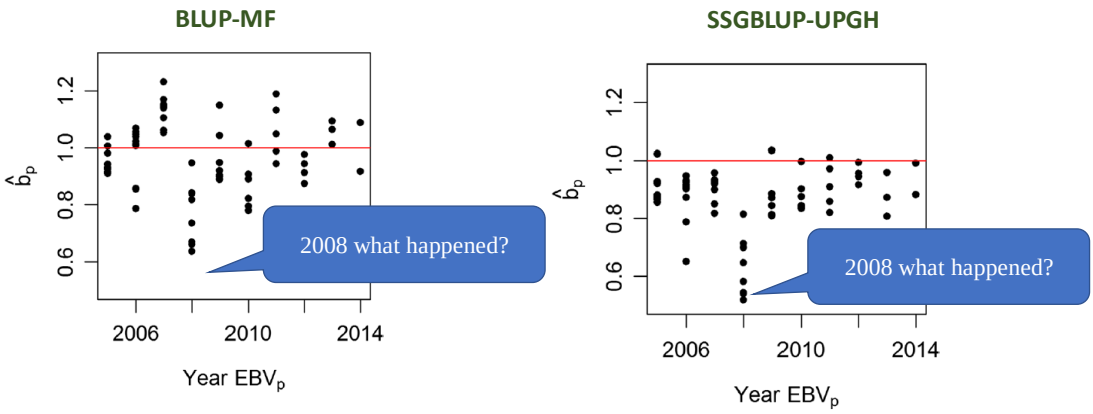
Very small bias in Manech Tête Rousse!
slope $\hat{b}_p$~1
small, positive bias $\hat{\Delta}_p$ (0.2 genetic s.d.)
lambs are over-predicted

Model	$\hat{\Delta}_p$	$\hat{b}_p$
BLUP-MF	0.25	0.98
BLUP-UPGA	0.48	0.96
SSGBLUP-MF	0.23	0.97
SSGBLUP-UPGA	0.32	0.94
SSGBLUP-UPGH	0.48	0.88

Across models
slope $\hat{b}_p$:
MF performs better

	EBV_w		
EBV_p	SSGBLUP-UPGA	SSGBLUP-UPGH	SSGBLUP-MF
BLUP-MF	1.32	1.29	0.98
BLUP-UPGA	1.25	1.23	0.92

Main results Manech Tête Rousse



Important variation between truncation points!

Examples of estimation of accuracies (MTR)

Model	Selected reliability $\widehat{acc}_p^2$	Unselected reliability $\widehat{rel}_p$	Ratio of reliabilities $\hat{\rho}_{pw}^2$
BLUP-MF	0.22	0.53	0.32
BLUP-UPG	0.24	0.54	0.31
SSGBLUP-MF	0.32	0.59	0.45

All of them agree in saying SSGBLUP >> BLUP

The "unselected reliability" is in the scale of Reliability that we are familiar with
The "ratio of reliabilities" is harder to interpret

Macedo et al. Genet Sel Eval (2020) 52:47
https://doi.org/10.1186/s12711-020-00567-1

RESEARCH ARTICLE Open A

Bias and accuracy of dairy sheep evaluations using BLUP and SSGBLUP with metafounders and unknown parent groups

REMOVING DATA AND USING METAFOUNDERS ALLEVIATES BIASES FOR ALL TRAITS IN LACAUNE DAIRY SHEEP PREDICTIONS

Journal:	*Journal of Dairy Science*
Manuscript ID	JDS.2021-20860.R1
Article Type:	Research
Date Submitted by the Author:	n/a
Complete List of Authors:	Macedo, Fernando; Institut National de la Recherche Agronomique; Universidad de la Republica, Facultad de Veterinaria; Swedish University of Agricultural Sciences, Animal Breeding and Genetics Astruc, Jean-Michel; Institut de l'Elevage (IDELE) Meuwissen, Theo; Agricultural University Norway, Inst. Animal Science Legarra, Andrés; INRAE, GenPhySE

SCENARIOS

Lacaune

Several scenarios:

OFFICIAL – Production traits and SCS as single trait; udder morphology as multiple trait

SINGLETRAIT – All evaluations are single trait

DELETION – Like "official" but deleting data (pedigree + records) since 1990

MULTIPLETRAIT – All traits in a multiple trait evaluation

Main results Lacaune: BIAS

Scenario	Model	Traits						
		MY	FC	PC	SCS	TA	UC	UD
DELETION	BLUP-UPGA	-0.02	-0.14	-0.22	-0.05	0.01	-0.06	-0.03
	SSGBLUP-MF1	-0.01	-0.09	-0.12	-0.03	0.01	-0.05	-0.02
	SSGBLUP-UPGH	-0.01	-0.09	-0.12	-0.03	0.01	-0.04	-0.03
OFFICIAL	BLUP-UPGA	0.15	-0.11	-0.20	-0.10	0.08	-0.12	-0.07
	SSGBLUP-MF1	0.11	-0.09	-0.14	-0.09	0.06	-0.10	-0.07
	SSGBLUP-UPGH	0.14	-0.08	-0.13	-0.08	0.07	-0.10	-0.06

Bias (overestimation of genetic trend) on "official" evaluations
Similar biases for SINGLETRAIT and MULTIPLETRAIT

The DELETION of historical data decreases the bias in almost all traits

Main results Lacaune: SLOPE

Scenario	Model	Traits						
		MY	FC	PC	SCS	TA	UC	UD
DELETION	BLUP-UPGA	0.96	1.01	0.96	0.96	0.99	0.97	0.92
	SSGBLUP-MF1	0.99	0.99	0.98	0.99	0.97	0.96	0.91
	SSGBLUP-UPGH	0.98	0.98	0.96	0.99	0.97	0.96	0.91
OFFICIAL	BLUP-UPGA	0.86	0.95	0.94	0.88	0.85	0.80	0.66
	SSGBLUP-MF1	0.87	0.93	0.91	0.87	0.83	0.79	0.73
	SSGBLUP-UPGH	0.80	0.94	0.92	0.87	0.83	0.79	0.73

DELETION improves the values of Slope

MF tend to perform better in Milk Yield OFFICIAL

Main results Lacaune: $\hat{\rho}_{p,w}$

Scenario	Model	Traits						
		MY	FC	PC	SCS	TA	UC	UD
OFFICIAL	BLUP-UPGA	0.45	0.57	0.59	0.52	0.69	0.75	0.61
	SSGBLUP-MF1	0.65	0.72	0.73	0.71	0.68	0.66	0.62

Almost all traits benefit from genomic prediction

Some do not – not clear why

Main conclusions of Dairy Sheep studies

- **There is an important variation in estimates of bias across different truncation points**
- **The deletion of historical data eliminates the bias in genetic evaluations without affecting the ranking of individuals**
- **In both works, the use of metafounders to manage missing pedigree performs better than (fixed) unknown parent groups**
- **The use of genomic information reduces bias and increases accuracy of the EBVs at birth**

What if my model is already wrong?

- The LR theory assumes that the model is correct !!
- Can we verify if a model is correct if the model is not correct?
- FM explored that

By simulation

- Software: QMsim, Blupf90 family and our own
- 20 replicates of a "dairy" population
- 10 generations
- Two heritabilities (0.1 and 0.3)
- Three scenarios

Correct Model

Genetic evaluations performed with correct heritabilities and effects

Wrong Heritability

Using higher (+0.05) and lower (-0.05) heritabilities in the evaluation model

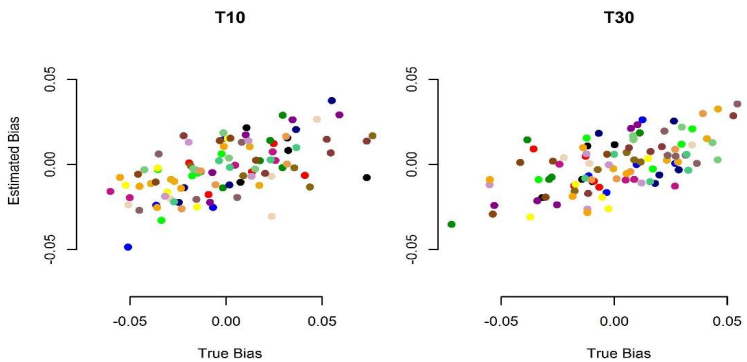
Environmental trend not (well) accounted for

Simulate a environmental trend.
Fit contemporary groups either as fixed, or as random heavily shrunken to 0 .

LR method

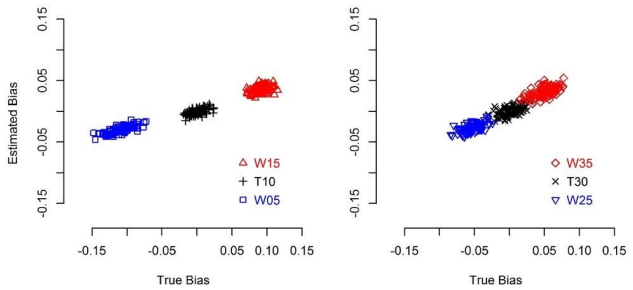
```
For generation 5 to 10{
    Compare males' EBVs at birth with the EBVs at "next" evaluation with daughter information.
    Get bias, slope and accuracies.
}
```


Main Results: the correct model



There was no surprise with the right model. Bias, slope, and accuracies were well estimated.

Main Results: the wrong heritabilities



The slope was estimated but with low precision

Accuracies were well estimated

True bias was generated
LR method could estimate the good direction but not the magnitude

Main Results: the environmental trend

It was impossible to estimate the Bias, neither fitting CG as fixed nor as random effect.

The slope was poorly estimated. The estimation was better when CG were fit as fixed effects.

In general, accuracies were well estimated

Main conclusion

The LR method can estimate the Bias, Slope, and Accuracies when the genetic evaluation model is robust, even if not perfect.

When the model is really wrong, the estimates from LR method are unreliable.

Take home messages

- In dairy sheep:
  - modelling of unknown parent groups in SSGBLUP is tricky and our best solution is metafounders
  - Lacaune has more bias than Manech Tete Rousse
    - possibly because it has more years of data and modelling is complicated
  - deleting old data is a very simple solution !!
    - even when the models are wrong, deleting old data alleviates the problem

Take home messages

- The LR methods gives a “very automatic” manner
  - of verifying that evaluations are correct
  - and of assessing accuracies empirically
- The most difficult part is to define the “focal groups” and to manipulate the data
  - you need to know the selection scheme and the data set
  - an unguided PhD student can not do it correctly
  - someone who doesn’t like scripting can not do it correctly
- It is very important to analyze multiple truncation points

Technical details and relevant literature

- Legarra, A., & Reverter, A. (2017, July). Can we frame and understand cross-validation results in animal breeding. In *Proceedings of the 22nd conference association for the advancement of animal breeding and genetics* (pp. 2-5).
- Legarra, A., & Reverter, A. (2018). Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genetics Selection Evolution*, *50*(1), 1-18.
- Legarra, A., & Reverter, A. (2019). Correction to: Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genetics Selection Evolution*, *51*(1), 1-2.
- Macedo, F. L., Reverter, A., & Legarra, A. (2020). Behavior of the Linear Regression method to estimate bias and accuracies with correct and incorrect genetic evaluation models. *Journal of dairy science*, *103*(1), 529-544.
- Macedo, F. L., Christensen, O. F., Astruc, J. M., Aguilar, I., Masuda, Y., & Legarra, A. (2020). Bias and accuracy of dairy sheep evaluations using BLUP and SSGBLUP with metafounders and unknown parent groups. *Genetics Selection Evolution*, *52*(1), 1-10.
- Bermann, M., Legarra, A., Hollifield, M. K., Masuda, Y., Lourenco, D., & Misztal, I. (2021). Validation of single-step GBLUP genomic predictions from threshold models using the linear regression method: An application in chicken mortality. *Journal of Animal Breeding and Genetics*, *138*(1), 4-13.

- A tutorial: http://genoweb.toulouse.inra.fr/~alegarra/GSIP.pdf (chapter 15)

Background

Development and theory

Simulation: it works

All the correct expressions + testing on MTR

More development + testing on chicken

Relevant old literature

- Mantysaari EA, Liu Z, VanRaden P. Interbull validation test for genomic evaluations. Interbull Bull. 2010;41:17–21.
- Thompson R. Statistical validation of genetic models. Livest Prod Sci. 2001;72:129–34.
- Reverter A, Golden BL, Bourdon RM, Brinks JS. Technical note: detection of bias in genetic predictions. J Anim Sci. 1994;72:34–7.
- Bijma P. Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations. J Anim Breed Genet. 2012;129:345–58.
- Dekkers JCM. Asymptotic response to selection on best linear unbiased predictors of breeding values. Anim Sci. 1992;54:351–60.
- Henderson CR. Best linear unbiased prediction in populations that have undergone selection. In: Proceedings of the world congress on sheep and beef cattle breeding: 28 October–13 November 1980; Palmerston North and Christchurch. Palmerston North: Dunmore Press; 1982. p. 191–201.
- Henderson CR. Best linear unbiased estimation and prediction under a selection model. Biometrics. 1975;31:423–47.