Genomic selection

1- Linkage disequilibrium

Andrés Legarra - INRA

Acknowledgements

- ANR projects Amasgen, Rules&Tools; Apisgene
- Toulouse bioinformatics platform (bioinfo.genotoul.fr)
- GENOMIA funding:
  www.poctefa.eu

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)

3

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

4

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

5

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**

6

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

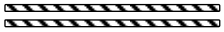
8

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

9

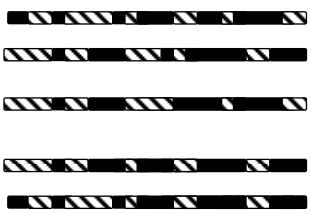
Populational linkage disequilibrium

- 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?

15

Populational linkage disequilibrium

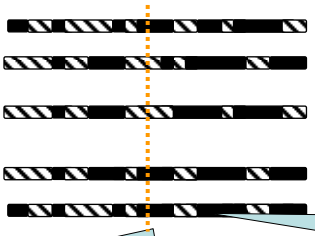
- The chromosomes become a fine-grained mosaic of grey and black



- However, complete mixture is difficult to attain

16

Populational linkage disequilibrium



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

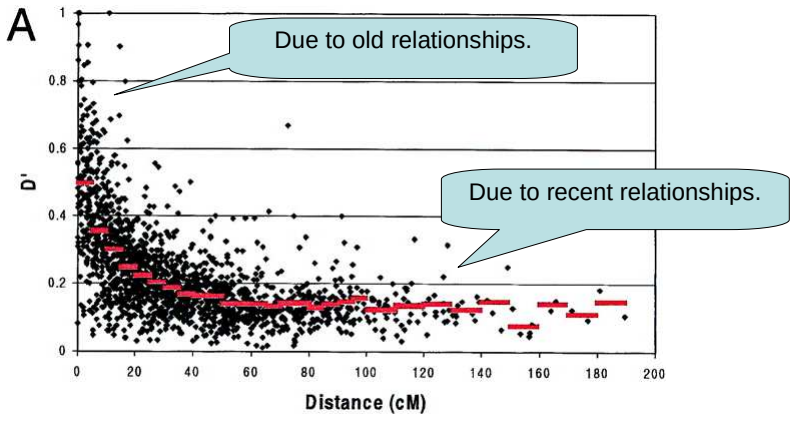
An stretch (=chromosomal segment) is conserved because it comes from the same ancestor (co-ancestry).

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

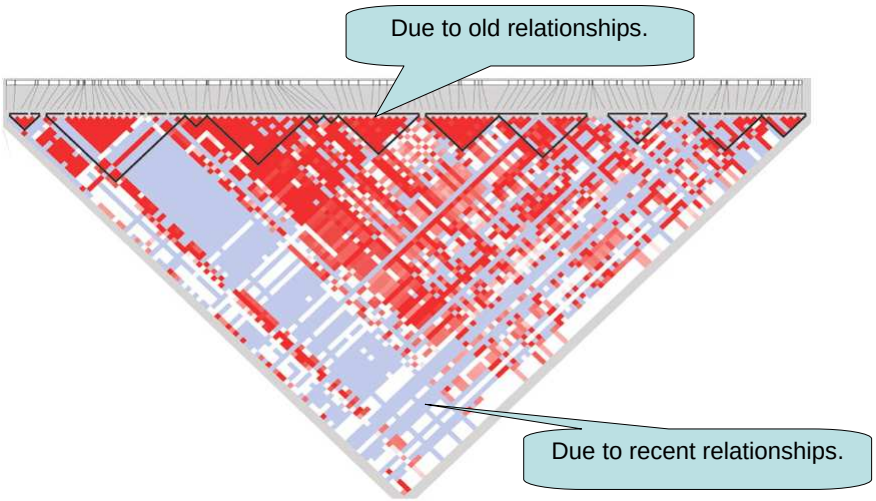
LD is therefore:
an <u>over-representation</u> of segments from a few gametes that existed in the population some time ago.

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

17



18



19

Populational linkage disequilibrium

- Mechanisms generating linkage disequilibrium of this kind are:
  - Drift
  - Selection
  - Migration
    - These three because at some point the chromosome of one animal is overrepresented
  - Mutation
    - considered by coalescence theory
    - less important in animal breeding

21

Genomic selection

2- Models for Genomic selection

Andrés Legarra - INRA

22

Models for Genomic selection

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

25

Single marker

- Assume there is a marker in complete LD with a QTL
- For example, the polymorphism in the halothane gene (HAL) which is a predictor of bad meat quality in swine

26

Single marker

- Estimate breeding values including the marker is a piece of cake
- $\mathbf{y}_i$= marker effect in animal *i* + *e*
  - We substitute the true, possibly unknown gene by a proxy observed marker and estimate effects of the latter using a linear model
  - We can include an additional polygenic genetic value of animal *i*

27

Base model

- **y= …+ Za + e**
  - **Z**= incidence matrix of marker effects
  - **a**= marker effect
  - **e**=residuals

3 individuals, 1 marker with 4 alleles

$$\mathbf{Za} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix}$$

- This can be solved, for example, by least squares

28

Single marker

- This is fine if we know what markers are good predictor of what genes
- But this is rarely the case
  - It can be shown that you miss a lot of information by trying to locate the QTLs
  - And those that you find are certainly exaggerated

29

Genomic selection

- So this is why we need genomic selection
- Although the idea is old, Meuwissen et al (2001) showed that high accuracies can be obtained with high density markers (under quite ideal scenarios)

34

Meuwissen's points

1. QTL mapping is a lot of work
   1. lots of people
   2. lots of time
   3. Not-automatic procedure
2. QTL detection is biased: estimated effect of « accepted » QTLs is exaggerated.
   1. Beavis effect
3. Very serious problem if many QTL positions are tested

35

Beavis effect

- We are mapping QTLs
- To declare a QTL in a position, we perform a test (for example a t-test)
- This test depends on the estimated effect of the QTL
  - estimated effect = real effect + « estimation noise »
  - by keeping selected QTLs, we often keep large and positive noises
    - this is negligible if there were few QTLs with large effects but this is not the case
    - large noises will occur in analysis with *many* markers
  - this biases the estimated QTL effect

36

Whole genome

- If we don't select QTL regions we skip the problem of bias
- Therefore :
  - Genetic value = sum of effects of all regions
    - We effectively treat all regions as being carriers of a QTL
  - How do we estimate the effects of all regions?

38

Whole genome

- The simpler is to do an extension of single marker analysis
- Do multiple marker regression
- Works well only with dense markers!

39

Multiple marker additive model

- **y**= **Za** + **e**
  - **Z**= incidence matrix of marker effects
  - **a**= marker effect
  - **e**=residuals

4 individuals, 2 markers each

2 alleles in 1st marker

$$\mathbf{Za} = \left[\begin{array}{cc|cccc} 1 & 1 & 0 & 1 & 1 & 0 \\ 2 & 0 & 2 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & 0 & 1 \\ 0 & 2 & 1 & 1 & 0 & 0 \end{array}\right] \left[\begin{array}{c} a_{1,1} \\ a_{1,2} \\ \hline a_{2,1} \\ a_{2,2} \\ a_{2,3} \\ a_{2,4} \end{array}\right]$$

4 alleles in 2nd marker

40

Estimating SNP effects

- The simultaneous estimates of many markers by least squares are very poor,
  - if we have much more SNPs than individuals
  - They are thus terribly bad for genomic predictions as well
- Even if we had many individuals, there is a missing piece of information:
  - we think that most SNPs do *not* have an effect or at least a big one
  - this is a « prior » information
- Can we do something?

42

Integration

- Need to postulate p(**a**) : this means SNP are <u>random</u> effects
- Best Predictor is optimal for selection (Goffinet and Elsen 1984 GSE; Gianola and Fernando 1986 JAS)
- The Best Predictor comes from the conditional mean of SNP effects after observing the data

$$\hat{\mathbf{a}} = E(\mathbf{a} \mid \mathbf{y}) = \frac{\int \mathbf{a}\, p(\mathbf{y} \mid \mathbf{a})\, p(\mathbf{a})\, d\mathbf{a}}{\int p(\mathbf{y} \mid \mathbf{a})\, p(\mathbf{a})\, d\mathbf{a}}$$

- Remember: classical BLUP is an example of BP if
  - $p(\mathbf{g}) \sim N(0, \mathbf{A}\sigma^2_u)$

43

A priori Distributions for marker effects

- *Unknown*
- *Still we don't have a good theory*
- Agreement on
  - Small effects more frequent than big ones
  - marker effects are assumed *a priori* independent

44

A priori Distributions for marker effects

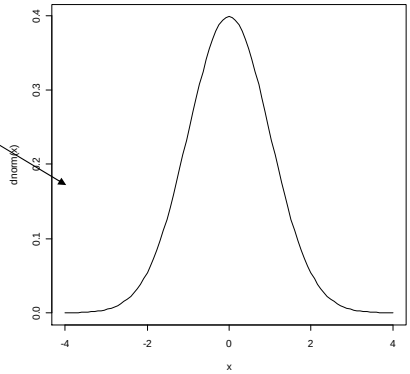
- Several distributions for SNP effects have been proposed
  - Normal (Meuwissen et al., Genetics 2001; Van Raden JDS 2008) -> BLUP_SNP or GBLUP or RR-BLUP
  - BayesA, BayesB, (Meuwissen et al. 2001; Habier et al., 2011)
  - Mixture of normal , BayesC(Pi) (Van Raden JDS 2008, Habier et al., 2011)
  - (Bayesian) Lasso (Usai et al., 2009; De los Campos, et al., 2009)

45

Normal distribution

$$a_i \sim N\left(0, \sigma_a^2\right)$$

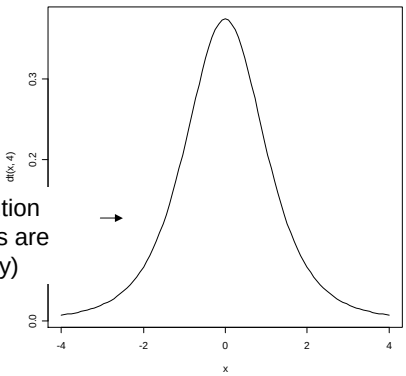
Few « big » effects



46

BayesA (*t* distribution)

a *t* distribution (big effects are not unlikely)



representation as « t »

$$a_i \sim t\left(0, \nu, \sigma_a^2\right)$$

$$\equiv$$

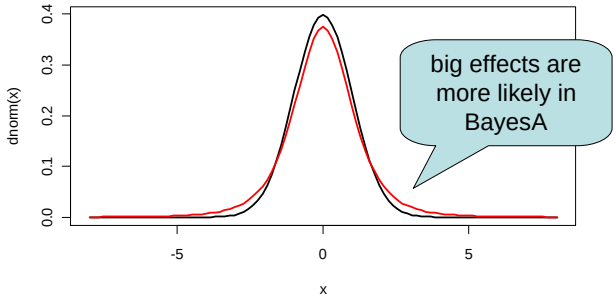
$$a_i \sim N\left(0, \sigma_{a,i}^2\right)\chi_\nu^{-2}\sigma_a^{-2}$$

Meuwissen et al. representation

Gianola et al. (2009) proved that fitting a variance by locus is equivalent to postulating t distribution for all locus

47

Normal vs. BayesA

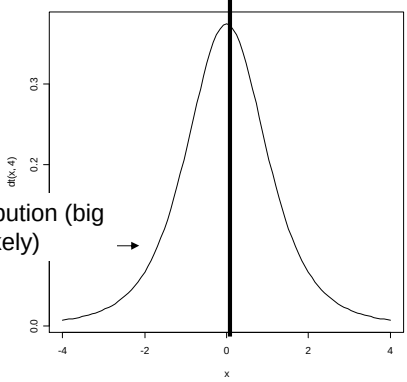


48

BayesB (mixture with *t* distribution)

$$a_i \sim \pi(0) + (1-\pi)\, t\left(0, \nu, \sigma_a^2\right)$$

A fraction π of markers has null effects.



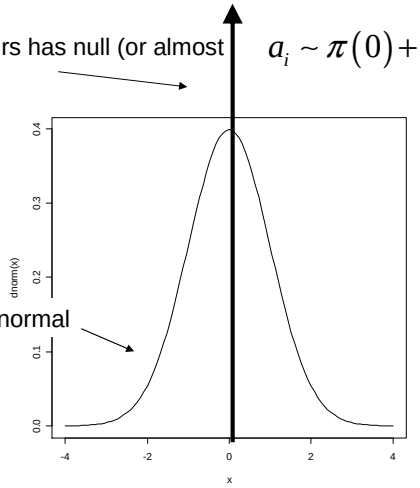
Otherwise a *t* distribution (big effects are not unlikely)

49

Mixture distribution or BayesC(Pi)

A fraction π of markers has null (or almost null) effects.

$$a_i \sim \pi(0) + (1-\pi)\, N\left(0, \sigma_a^2\right)$$



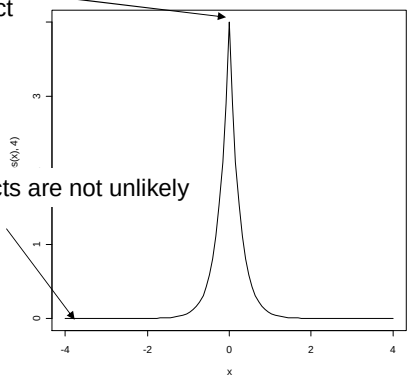
Otherwise they are normal

50

Lasso (double exponential)

$$a_i \sim \frac{\lambda}{2}\exp\left(-\lambda\left|a_i\right|\right)$$

Often marker has *almost* null effect



Otherwise big effects are not unlikely

51

What is the best distribution?

- Depends on what truth is ☺
- Not a good quantitative genetics theory to choose one
- Funny distributions (mixtures, BayesB, double exponential) have tricky parameters (e.g. π).
- Optimal methods are not the same for different data (traits)
- Normal distribution quite good

52

Recipe for genomic selection

- Genotype all your animals
- Get records (e.g. weight)
- Pick a model for **a**, say normal distribution
- **y=Xb + Za + e**
  - **b**: fixed effects (herd, year)
  - **a**: SNP effects
- Solve for **â** using some method

53

Recipe for genomic selection

- What do you do with a candidate for selection – say a young bull – with no data or daughters?
  - Genotype it
  - build its incidence matrix for SNPs $\mathbf{z}_i$
  - take the estimates for SNP effects $\hat{\mathbf{a}}$
  - The estimated breeding value is EBV=$\mathbf{z}_i\hat{\mathbf{a}}$

54