Opportunities and limits for the use of Equine SNP50 Bead chips

Bertrand Langlois
Bertrand.langlois@jouy.inra.fr
SGQA-INRA 78350 Jouy-en-Josas, France.
Eaap Barcelonna S19-8

INTRODUCTION-a

- Sufficient SNP were mapped to the Equ Cab 2.0 assembly to obtain a 50 000 SNP illumina array.
- This resulted in an Equine 50 Beadchip which is now commercially available.
- However the cost for genotyping is currently about 250 Euros per animal.
- The experimental design needed to answer different questions is therefore of great economical concern.

INTRODUCTION-b

- In this paper we will check these problems for horse population genetics.
- Three levels will be considered:
- -1- The estimation of allele frequencies allowing the calculation of genetic distances between breeds or individuals (kinship with markers)
- -2- Detecting signatures of selection for those loci not significantly in Hardy Weinberg equilibrium.
- -3- The possibility to check for differences in linkage disequilibrium intra and between breeds.

1-a The estimation of allele frequencies

- The confidence interval x of a frequency F is generally written

$$x = \frac{r}{n} \pm 1{,}96 \sqrt{\frac{F(1-F)}{N}}$$

- r is the number of realisations of the event
- N is the number of sorting
- F(1-F) is maximum for F=0.5
- Therefore if you want to know F with 1% precision at risk 5%

- 1,96 x 0.5 = 0.01 $N^{1/2}$ lead to N=98

- At risk 1% you have to change 1.96 by 2.576 which lead to N= 129

1-b The estimation of allele frequencies

- Knowing that every individual is wearing a couple of genes
- For 5 & 1% confidence interval for the allele frequency
- you need genotyping between 50 and 65 individuals

2-a Detecting signatures of selection

- The estimation of the departure from panmixia for allele i

$$\hat{D}_{ii} = \hat{F}_{ii} - \hat{p}_i^2$$

- This estimation resulting of N observations is biaised and Weir (1996) wrote

$$Esp.(\hat{D}_{ii}) = D_{ii} - \frac{1}{2N}(p_i + F_{ii} - 2p_i^2)$$

$$= D_{ii} - \frac{1}{2N}[p_i(1 - p_i) + D_{ii}]$$

- The factor of correction is rapidely tending to zero when N>50

2-b Detecting signatures of selection

- According to Weir 1996 (applying Fisher's approximation)

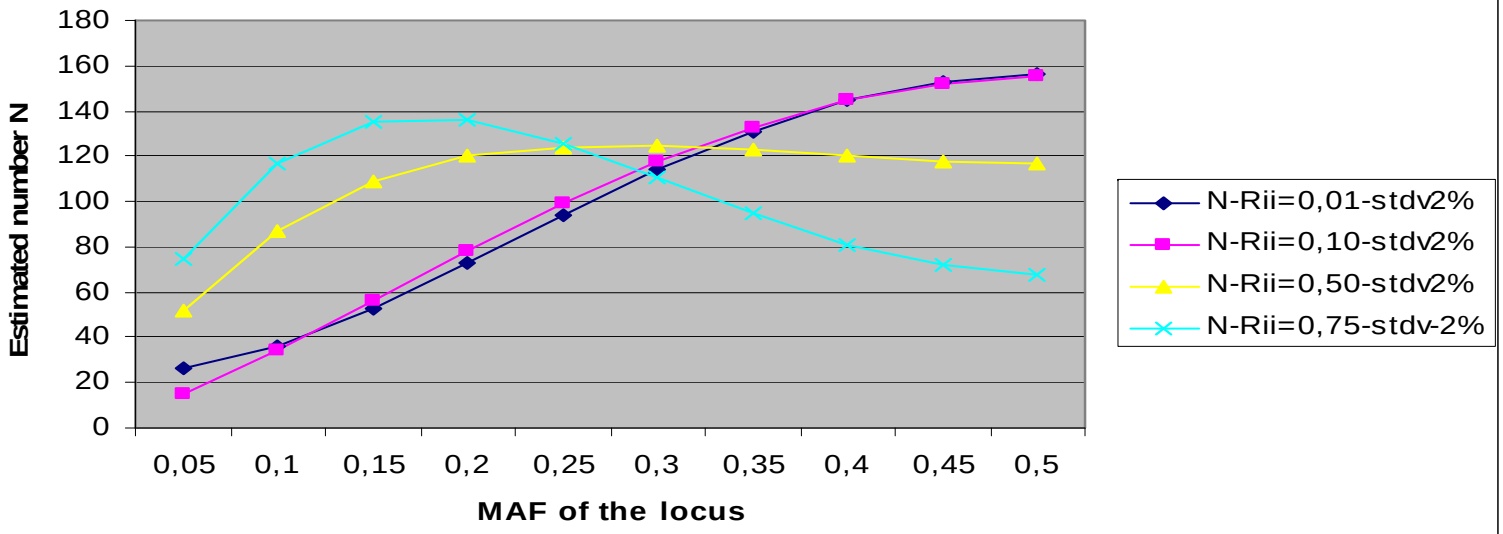
$$Var(\hat{D}_{ii}) \approx \frac{1}{N}\left[p_i^2 q_i^2 + (1 - 2p_i)^2 D_{ii} - D_{ii}^2\right]$$

- If we want a standard déviation of 1% that is variance = 0.0001 it is possible to calculate N for different values of the alleles frequencies and the *a priori* on the disequilibrium $D_{ii} = R_{ii}\ p_i q_i$

2-c Detecting signatures of selection

- Stdv o.o1 needs four times the number needed for stdv 0.02. We present therefore only one figure for the two cases (note 125<N<150 for stdv 2% leads to 500-600 for stdv 1%)

Figure 1:Necessary N for estimating Dii the deviation from panmixia with a standard deviation of 2% for différent a priori values of Rii from WEIR (1996)



3-a The possibility to check for differences in linkage disequilibrium

- The estimation of the departure from linkage equilibrium for alleles a (the most frequent) at loci i and j

$$\hat{D}_{ij} = \hat{F}_{ij} - \hat{p}_i\hat{p}_j$$

- Is also biaised

$$Esp.(\hat{D}_{ij}) = \hat{F}_{ij} - \hat{p}_i\hat{p}_j$$

$$Esp\,(\hat{D}_{ij}) = \frac{2N^* - 1}{2N^*} D_{ij}$$

N* concern the alleles ai aj. Four N* can be defined (N_1*+N^*_2+N_3*+N_4*=N). When N* >50 meaning N >200 the bias is neglectable.

3-b The possibility to check for differences in linkage disequilibrium

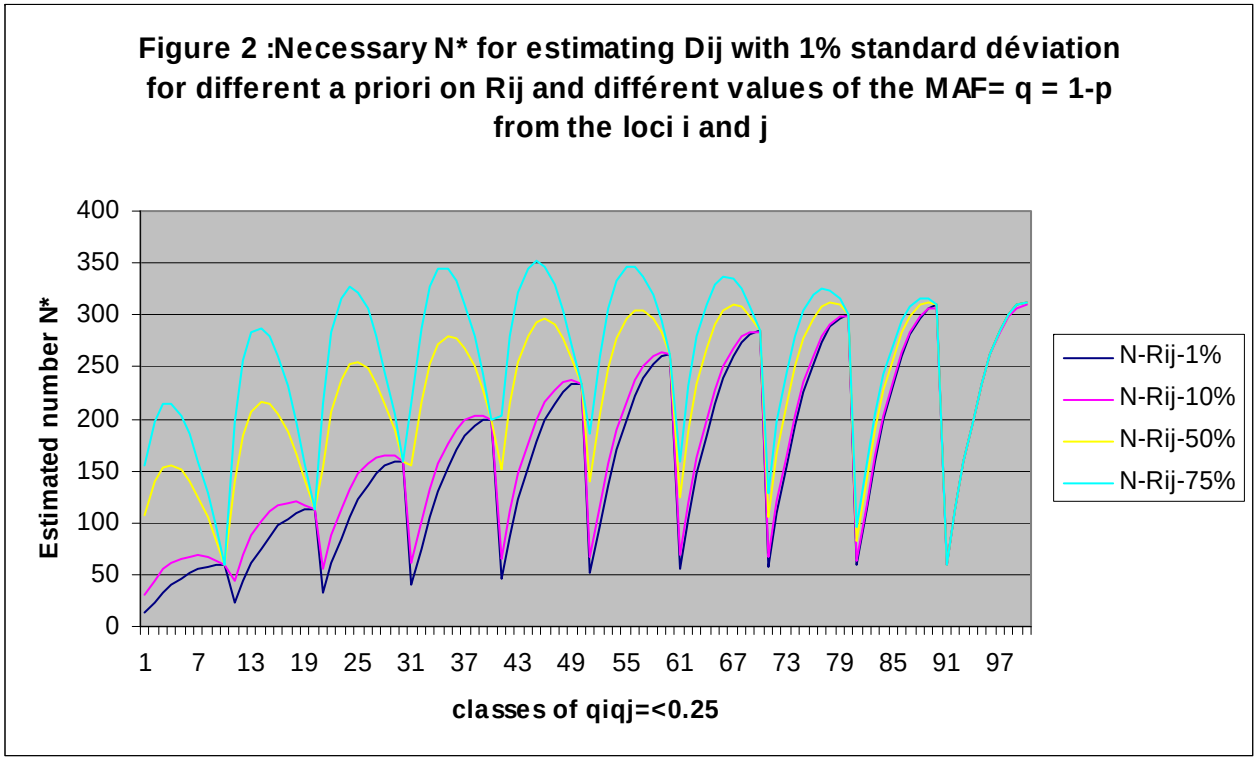
- According to Weir (1996) (applying Fisher's approximation)

$$Var\left(\hat{D}_{ij}\right) = \frac{1}{2N^*}\left[p_i q_i p_j q_j + (1 - 2p_i)(1 - 2p_j)D_{ij} - D_{ij}^2\right]$$

- If we want a standard deviation of 1% that is variance = 0.0001 it is possible to calculate N* for different values of the alleles frequencies
- and the *a priori* on the disequilibrium

$$Dij = Rij\ (piqipjqj)^{1/2}$$

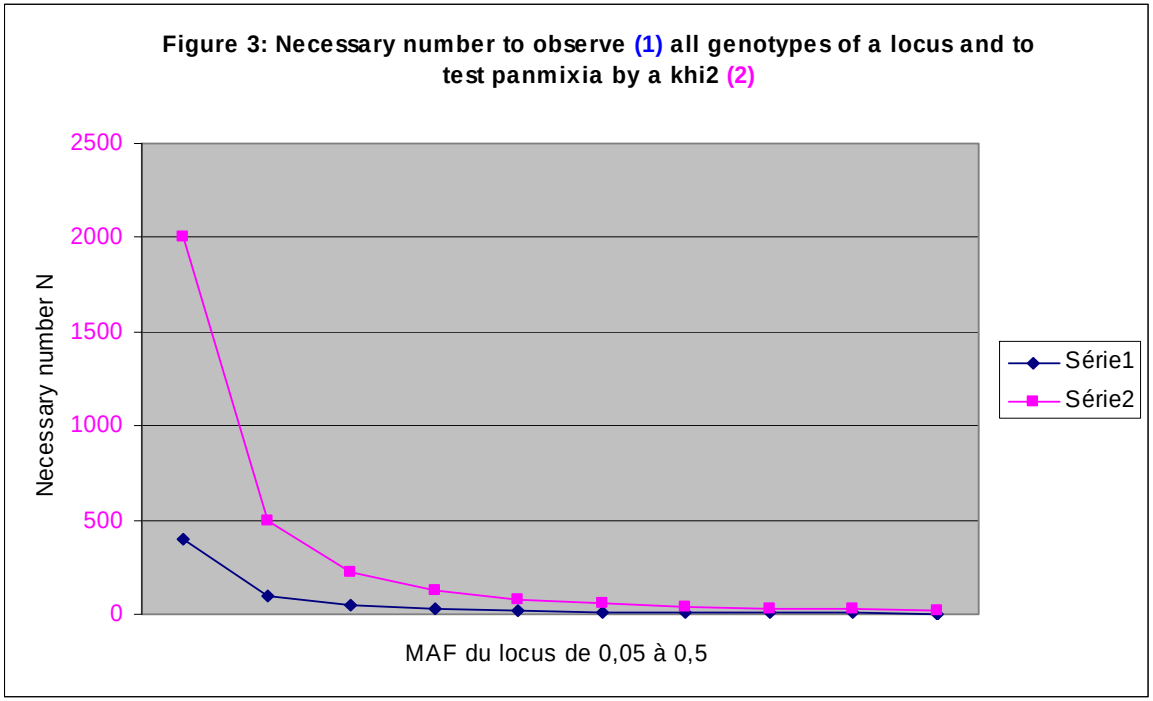
3-c The possibility to check for differences in linkage disequilibrium



- We can also consider that N* stdv 1 % equals N stdv 2 %.
- (note that 300<N<350for 2% stdv leads to 1200<N<1400 for 1% stdv)

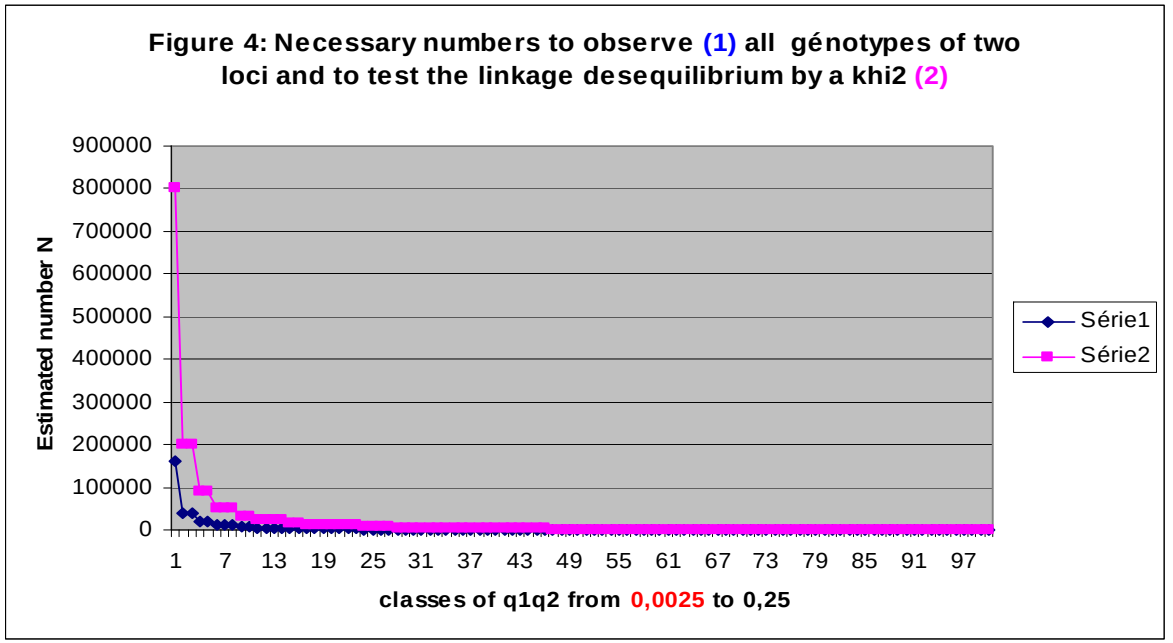
4-a Khi 2(df 3-2=1) testing panmixia

- The necessary number to observe and to test rare genotypes are respectively $N=1/q^2$ and $N= 5/q^2$

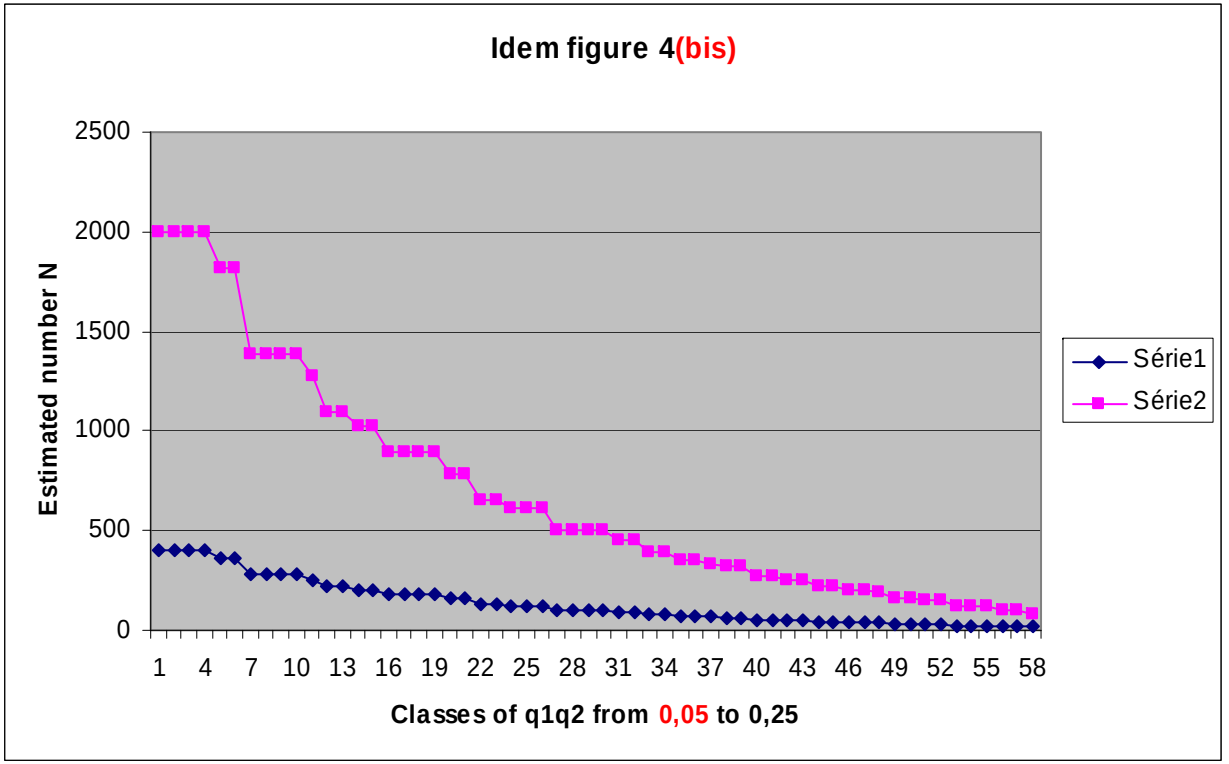


4-b Khi 2(df 9-3=6) for testing Linkage disequilibrium

- The necessary number to observe and to test rare genotypes (double rare homozygous) are respectively $N=1/q_1^2q_2^2$ and $N= 5/q_1^2q_2^2$



Focus for geometric mean of q_1 & q_2 >0.22and<0.5



SYNTHESIS

N		$(5/N)^{1/2}$ MAF necessary to test panmixia	$(5/N)^{1/4}$ Geometric mean of MAF necessary to test a linkage disequilibrium
50	(5%)	0,32	0,56
65	(1%)	0,28	0,53
125	Dii 2%	**0,20**	**0,45**
150	Dii 2%	0,18	0,43
300	Dij 2%	**0,13**	**0,36**
350	Dij 2%	0,12	0,35
500	Dii 1%	**0,10**	**0,32**
600	Dii 1%	0,09	0,30
1200	Dij 1%	**0,06**	**0,25**
1400	Dij 1%	0,06	0,24
2000		**0,05**	0,22
3000		0,04	**0,.20**

CONCLUSION

- The conclusion is that for purpose1, N # 50-65 is sufficient.
-
- However a minimum of N #125-150 is requested for purpose 2.
- But N#500-600 would be better

- N # 2000-3000 would be good for purpose 3, but 1200-1400 would be enough and 300-350 would allow a first approach

- Indeed, a low N allow approaching all three levels but only for SNP with high MAF values.

- Bibliography: WEIR B. 1996. Genetic Data Analysis II chapter 3 p95 & p113. Sinauer Associates,Publishers Sunderland, Massachusetts.