What Could Be Expected For QTL Detection From EquineSNP50 BeadChip And Simple SNP Analysis?

A. Ricard[*], S. Teyssèdre[†] and J.M. Elsen[†]

Introduction

The availability of EquineSNP50 bead-chip reinforced the opportunity of QTL research for horses in a wild range of traits, including health as well as performances. Primary results of molecular analysis were on hereditary diseases caused by single gene effects and coat color genetics (http://www.uky.edu/Ag/Horsemap/hgpapps.html). But, recently, publications involved performance traits (Hill *et al.* (2010)). It is too early to present a summary of literature as most of new results will certainly be presented in that meeting. Most popular methods used to detect QTL with SNP are based on linkage disequilibrium (LD). The aim of this study was to evaluate the robustness and power of simple single SNP test as regression test and GRAMMAR (Aulchenko *et al*. (2007)) in the practical case of equine population samples, i.e. with uncontrolled relationships and controlled paternal half sib's families of small size.

Material and methods

The objective was to found the mean and variance of the test statistics used in regression and GRAMMAR analysis under the assumption of a true model involving both SNP effect and additive genetic value. Then robustness and power were calculated for a sample of animals taken at random in one horse breed, so with relationships, and also for a sample built for linkage analysis so with half sibs families as usual in animal breeding.

True model. The true model supposed the co-existence of SNP effect and additive genetic value, as in the following mixed model:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{x}\beta + \mathbf{u} + \mathbf{e} \qquad (1)$$

where **y** was the vector of observed trait (one performance by horse), **μ** the vector of overall mean, β the regression coefficient of SNP effect, **u** the vector of additive genetic effects and **e** the vector of residuals. **x** was the incidence vector of SNP effect, defined as $\mathbf{w} - \bar{\mathbf{w}}$ where w was $-2p/\sqrt{2pq}$ for genotype 11, $(1-2p)/\sqrt{2pq}$ for genotype 12, $2q/\sqrt{2pq}$ for genotype 22, with p the frequency of allele 2. So that, assuming Hardy-Weinberg equilibrium, $E(w) = 0$ and $V(w) = 1$. Let $V(\mathbf{u}) = \mathbf{A}\sigma_u^2$ with **A** the relationship matrix and $h^2 = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$ the heritability of the trait. For simplicity no other fixed effect was added to the model.

[*]INRA, UMR 1313, 78352 Jouy-en-Josas, France
[†] INRA, UR 631, 31326 Castanet-Tolosan, France

Regression analysis. The classical regression analysis with one SNP at each time assumed a fixed model:

$$\mathbf{y} = \boldsymbol{\alpha} + \mathbf{x}\beta + \boldsymbol{\varepsilon} \quad (2)$$

With $\mathbf{y}$ and β as previously defined, $\boldsymbol{\alpha}$, the vector of mean and $\boldsymbol{\varepsilon}$ the vector of residuals.

GRAMMAR analysis GRAMMAR analysis is a two step model. First, genetic additive values were estimated with a random model:

$$\mathbf{y} = \boldsymbol{\delta} + \mathbf{a} + \boldsymbol{\gamma} \quad (3)$$

With $\begin{bmatrix} \mathbf{1'1} & \mathbf{1'} \\ \mathbf{1} & \mathbf{I} + \lambda\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\delta} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{1'y} \\ \mathbf{y} \end{bmatrix}$ where $\lambda = \frac{\sigma_\gamma^2}{\sigma_a^2}$ and $\begin{bmatrix} \hat{\delta} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{C_{11}} & \mathbf{C_{1a}} \\ \mathbf{C_{a1}} & \mathbf{C_{aa}} \end{bmatrix} \begin{bmatrix} \mathbf{1'y} \\ \mathbf{y} \end{bmatrix}$. Then a second fixed model was applied on residuals:

$$\mathbf{y} - \hat{\mathbf{a}} - \hat{\boldsymbol{\delta}} = \boldsymbol{\eta} + \mathbf{x}\beta + \boldsymbol{\varepsilon} \quad (4)$$

True distribution of test statistic in regression analysis. SNP effect was estimated by $\hat{\beta} = (\mathbf{x'x})^{-}\mathbf{x'y}$. Assuming the true model (1) and using the regression model (2) the distribution of $\hat{\beta}$ was:

$$V(\hat{\beta}) = \mathbf{(x'x)^{-}x'}V(\mathbf{y})\mathbf{x(x'x)^{-}} = \mathbf{(x'x)^{-}x'}(\mathbf{A}\sigma_u^2 + \mathbf{I}\sigma_e^2)\mathbf{x(x'x)^{-}} = \mathbf{(x'x)^{-}x'Ax(x'x)^{-}}\sigma_u^2 + \mathbf{(x'x)^{-}}\sigma_e^2$$

$$E(\hat{\beta}) = \mathbf{(x'x)^{-}x'}E(\mathbf{y}) = \mathbf{(x'x)^{-}x'}(\mathbf{x}\boldsymbol{\beta} + \bar{\mathbf{y}}) = \beta$$

$$E(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}) = (n-2)\sigma_e^2 + (tr(\mathbf{A}) - \mathbf{(x'x)^{-}x'Ax} - \frac{1}{n}\mathbf{1'A1})\sigma_u^2$$

So the test statistic used $\hat{\beta}\sqrt{\frac{\mathbf{x'x}(n-2)}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}}$ will not follow exactly a student distribution as residuals were no longer independent but will be closed to it, and close to a normal distribution when *n*, the total number of horses with genotypes and performances, is sufficiently high (>150). The mean of this normal distribution was $M = \beta\sqrt{\frac{\mathbf{x'x}(n-2)}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}}$ and the variance $V = \dfrac{(1-h^2) + \mathbf{x'Ax(x'x)^{-}}h^2}{(1-h^2) + \dfrac{h^2}{n-2}(tr(\mathbf{A}) - 1'\mathbf{A}1/n - \mathbf{x'Ax(x'x)^{-}})}$

True distribution of test statistic in GRAMMAR analysis. Assuming the true model (1), and using the random model (3) the expectation and variance of estimates of genetic values were:

$$E(\hat{\mathbf{a}}) = (\mathbf{C_{a1}1'} + \mathbf{C_{aa}})E(\mathbf{y}) = \mathbf{C_{aa}x}\beta$$

$$V(\hat{\mathbf{a}}) = (\mathbf{C_{a1}1'} + \mathbf{C_{aa}})V(\mathbf{y})(\mathbf{C_{a1}1'} + \mathbf{C_{aa}})' = \sigma_u^2(\mathbf{A} - \lambda\mathbf{C_{aa}}) + (\sigma_e^2 - \lambda\sigma_u^2)\mathbf{C_{aa}}(\mathbf{I} - \lambda\mathbf{A}^{-1}\mathbf{C_{aa}})$$

So, when applying the model (4) to estimate SNP effect $\hat{\beta} = \mathbf{(x'x)^{-}x'}(\mathbf{y} - \hat{\mathbf{a}} - \hat{\boldsymbol{\delta}})$,

$E(\hat{\beta}) = \mathbf{(x'x)^{-}x'}E(\mathbf{y} - \hat{\mathbf{a}} - \hat{\boldsymbol{\delta}}) = \beta - \mathbf{(x'x)^{-}x'C_{aa}x}\beta$, the test was biased.

$$V(\hat{\beta}) = \mathbf{(x'x)^{-}x'}V(\mathbf{y}-\hat{\mathbf{a}}-\mathbf{1}\hat{\delta})\mathbf{x(x'x)^{-}}$$
$$= \sigma_e^2\left[\mathbf{(x'x)^{-}} - \mathbf{(x'x)^{-}x'C_{aa}x(x'x)^{-}}\right] - (\sigma_e^2 - \lambda\sigma_u^2)\lambda\mathbf{(x'x)^{-}x'C_{aa}A^{-1}C_{aa}x(x'x)^{-}}$$

First, we assumed that bias and variance of estimation of λ used in model (3) was negligible (for example, estimate of λ was obtained in more general analysis) so that $(\sigma_e^2 - \lambda\sigma_u^2) = 0$. Then

$V(\hat{\beta}) = \sigma_e^2\left[\mathbf{(x'x)^{-}} - \mathbf{(x'x)^{-}x'C_{aa}x(x'x)^{-}}\right]$ and

$$E(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}) = \sigma_e^2\left[n-1-tr(\mathbf{C_{aa}}) + \mathbf{(x'x)^{-}x'C}_{aa}\mathbf{x}\right] + \beta\mathbf{x'C}_{aa}(\mathbf{I}-\mathbf{x(x'x)^{-}x'})\mathbf{C}_{aa}\mathbf{x}\beta$$

(without taken into account reduction of variance due to estimation of the mean δ)

So the test statistic $\hat{\beta}\sqrt{\dfrac{\mathbf{x'x}(n-2)}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}}$ had variance:

$$V = \frac{\sigma_e^2(n-2)(1-\mathbf{(x'x)^{-}x'C_{aa}x})}{\sigma_e^2\left[n-1-tr(\mathbf{C_{aa}}) + \mathbf{(x'x)^{-}x'C}_{aa}\mathbf{x}\right] + \beta\mathbf{x'C}_{aa}(\mathbf{I}-\mathbf{x(x'x)^{-}x'})\mathbf{C}_{aa}\mathbf{x}\beta}$$

and expectation $M = (\beta - \mathbf{(x'x)^{-}x'C_{aa}x}\beta)\sqrt{\dfrac{\mathbf{x'x}(n-2)}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}}$.

Expectation of mean and variance of the test statistic according to the distribution of SNP – Regression analysis. The objective was to validate the test according to the distribution of SNP, so relative to **x**. As $E(w_i w_j) = a_{ij}$ the relationship coefficient between the individuals *i* and *j*,

$$E_x(\mathbf{x'x}) = tr(\mathbf{A}) - \frac{1}{n}\mathbf{1'A1}$$

$$E_x(\mathbf{x'Ax}) = tr(\mathbf{A'A}) - \frac{2}{n}\mathbf{1'A'A1} + \frac{1}{n^2}\mathbf{1'A11'A1}$$

The relationship matrix may be divided in diagonal coefficient (D) and outside-diagonal coefficient (O). Let μ_D bet he mean of diagonal coefficients, μ_O the mean of outside diagonal coefficients and V_A the variance of all coefficients of the relationship matrix.

$E_x(\mathbf{x'x}) = (n-1)(\mu_D - \mu_O)$, $E_x(\mathbf{x'Ax}) \approx n^2 V_A$

Developing the variance of the coefficients of relationship matrix as:

$V_A = \dfrac{1}{n^2}\left[nV_D + n(n-1)V_O + (n-1)(\mu_D - \mu_O)^2\right]$ with V_D the variance of diagonal and V_O the variance of outside diagonal coefficients of relationship matrix, the expectation of the variance of the test statistic was:

$$E_x(V) = \frac{(1-h^2) + c_1 h^2}{(1-h^2) + c_2 h^2} \text{ with}$$

$$c_1 = (\mu_D - \mu_O) + \frac{n\left[V_D + (n-1)V_O\right]}{(n-1)(\mu_D - \mu_O)} \text{ and } c_2 = (\mu_D - \mu_O) - \frac{n\left[V_D + (n-1)V_O\right]}{(n-2)(n-1)(\mu_D - \mu_O)}$$

So that, it was the variability of the relationships between horses of the genotyped sample which modified the variance of the test and not the overall mean of relationships.

For half sibs families, $\mu_D = 1$, $\mu_O = \frac{n_d - 1}{4(n-1)}$ with $n_d = \frac{1}{n}\sum_{i=1}^{s} n_i^2$ with s the number of sires and n_i the number of progeny of the sire i (n_d is the number of progeny per sire for equal family size), $V_D = 0$, $V_O = \frac{(n_d - 1)(n - n_d)}{16(n-1)^2}$. By approximation, the variance in practical situation was the sum of variances due to families and due to background of relationship present in the horse breed studied.

Expectation of mean and variance of the test statistic according to the distribution of SNP – GRAMMAR analysis. In that case, in addition to $E_x(\mathbf{x'x})$, we had to calculate

$$E_x(\mathbf{x'C_{aa}x}) = tr(\mathbf{C_{aa}A}) + \frac{1}{n^2}\mathbf{1'A11'C_{aa}1} - \frac{2}{n}\mathbf{1'AC_{aa}1}$$

$$= tr(\mathbf{A}) - \lambda tr(\mathbf{C_{aa}}) - tr(\mathbf{C_{a1}1'A}) + \frac{1}{n^2}\mathbf{1'A11'C_{aa}1} - \frac{2}{n}\mathbf{1'AC_{aa}1}$$

$$\approx tr(\mathbf{A}) - \lambda tr(\mathbf{C_{aa}})$$

This was the sum of reliabilities of $\hat{a}$.

Similary $E_x(\mathbf{x'C_{aa}C_{aa}x}) \approx tr(\mathbf{A}) - \lambda tr(\mathbf{C_{aa}}) - \lambda tr(\mathbf{C_{aa}C_{aa}})$. For half sibs families, these reliabilities were equal to $\frac{h^2(n_i + 16 + 2h^2(n_i - 2))}{16 + 4(n_i - 1)h^2 - n_i h^4}$ with n_i number of progeny of the sire i.

To calculate $tr(\mathbf{C_{aa}C_{aa}})$ covariances between half sibs genetic values estimates were computed.

False discovery rate. The false discovery rate (FDR) was computed for a significance level α. Let $t_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ then $FDR = 2(1 - \Phi(\frac{t_{\alpha/2}}{\sqrt{E_x(V)}}))$ with Φ the cumulative normal distribution and $E_x(V)$ the expectation of variance of test statistic assuming null hypothesis.

Statistical power. The statistical power was computed as $P = 1 - \Phi(\frac{t_{\alpha/2} - E_x(M)}{\sqrt{E_x(V)}}))$, according a false discovery rate α, $E_x(M)$ and $E_x(V)$ are the expectation of mean and variance of test statistic under $\beta \neq 0$. The relation between the regression coefficient and the allele substitution effect (difference between genotype 11 and 12 or 12 and 22) was: $\beta_{allele} = \beta / \sqrt{2pq}$. So, the same statistical power is obtained for different substitution allele effects according to the frequency of the allele (MAF, minimum allele frequency). The power was also computed for a whole genome analysis according to EquineSNP50 Bead-Chip and linkage disequilibrium. The test for the QTL was supposed to be done with the closest marker. The power of this test depended on the correlation between QTL and SNP as $\beta_{SNP} = r_{QTL,SNP}\beta_{QTL}$ with $r_{QTL,SNP}$ the correlation between genotypes of SNP and QTL. The

distribution of correlations between adjacent markers was computed on sample of French Trotter population. We assumed that the linkage disequilibrium between putative QTL and closest SNP was the same than between adjacent SNP of the EquineSNP50 BeadChip. We summed over the distribution of these correlations the powers calculated for each correlation (by step 0.01).

Application. The formulae were applied to a sample size of 600 horses. Family structure was either random sample from horse French Trotters population or with half sibs family from 5 to 60 progeny per sires or both (families and relationship between families representative of trotter population). To calculate mean and variance of relationship matrix, the sample of Trotters was the 614 horses used in French program for osteochondrosis QTL detection. FDR was computed with corresponding 1% and 5% FDR for model without family structure The power was calculated for substitution allele effect 0.33 for MAF 10% equivalent to effect 0.20 with MAF 50% and effect 0.17 for MAF 10% equivalent to 0.10 for MAF 50%. The LD was measured on a sample of 267 genotyped French Trotter with EquineSNP50 BeadChip on 41249 markers selected from the 54602 SNPs with call freq>0.8, MAF>5% and P(test HardyWeinberg) $<10^{-8}$ (X chromosome excluded).

Results

Parental situation of French Horse Trotter. The mean and variance of relationship matrix diagonal and outside diagonal coefficients were: $\mu_D = 1.0415$, $\mu_O = 0.0553$, $V_D = 0.0004961$ $V_O = 0.0007970$ (paternal half sib relationships were removed from this distribution). The linkage disequilibrium between adjacent markers used in the calculation of total power is presented in Figure 1.

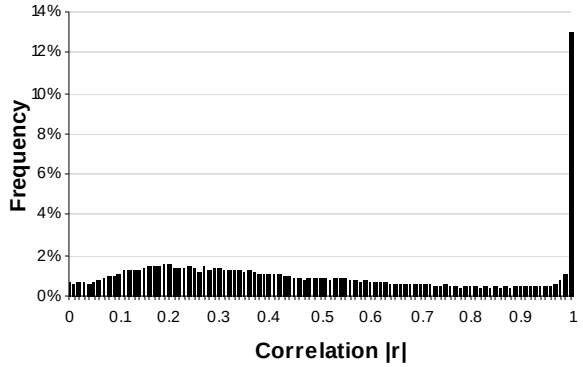

Figure 1: Distribution of Linkage Disequilibrium (|r|, absolute correlation between genotypes) between adjacent SNPs

Robustness. Figure 2 showed FDR expected according to family structure and a FDR of 5% without family structure for regression and GRAMMAR analysis respectively. With 1% of FDR without family structure, related populations with half sibs families (5 to 60 progeny by sire) gave FDR between 2% to 11% for h^2=0.40 and 4% to 21% for h^2=0.80. For GRAMMAR analysis, the FDR was 7‰ to 3‰ and 5‰ to 2‰ respectively.

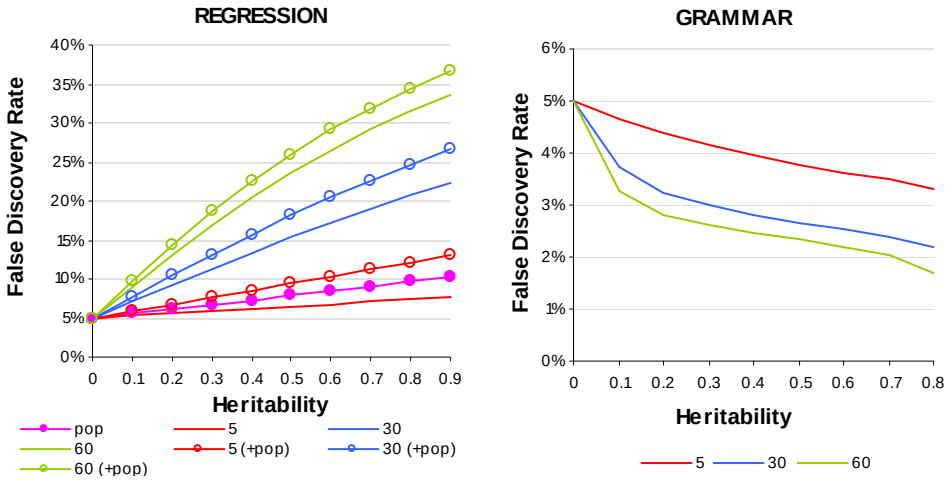

Figure 2: False discovery rate of regression and grammar analysis according to heritability and population structure: random sample from trotter population (pop) or with half sibs families (5/30/60 progeny) or both

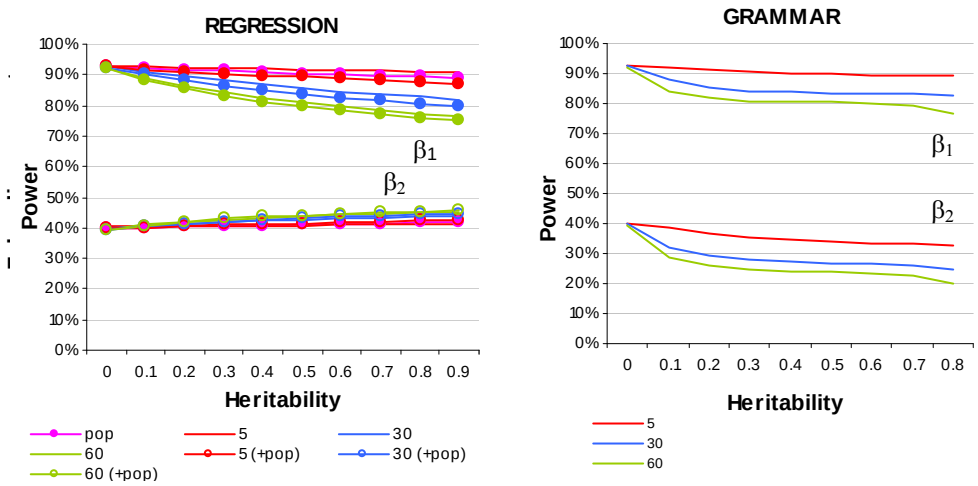

Figure 3: Power according to heritability and population structure for a substitution allele effect of 0.33 (MAF 10%) or 0.20 (MAF 50%) (β_1) and 0.17 (MAF 10%) or 0.10 (MAF 50%) (β_2) and FDR 5% when no family structure

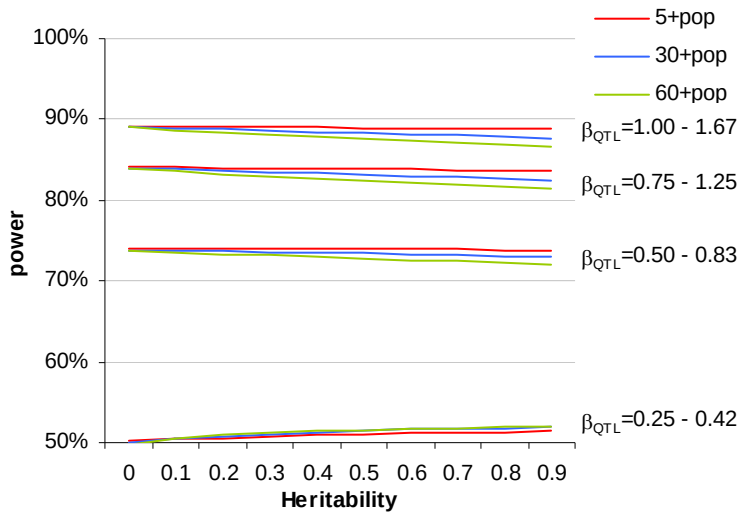

Figure 4. Power of the regression analysis according to heritability and population structure (3 family sizes with also population relationships) over the whole genome and QTL effects for MAF 50% and 10%.(first and second value)

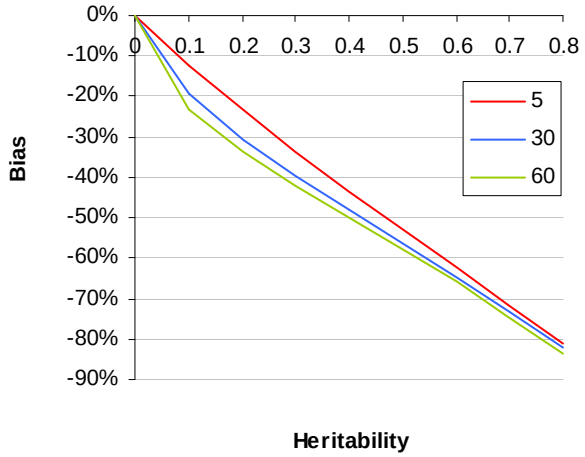

Figure 5: Bias of estimate of SNP effect $(E(\hat{\beta})-\beta)/\beta$ with GRAMMAR analysis according to heritability and population structure (from 5 to 60 progeny by sire)

Statistical power. For regression analysis, the power increased or decreased with increase of population structure according to the true regression coefficient (Figure 3). This is due to the position of the mean of the test relative to the threshold $t_{\alpha/2}$ and the increase of the variance. For grammar analysis the power decreased with increase of relationships because the variance of the test decreased. The power over the genome is presented in Figure 4. With

FDR 1%, the power decreased from this figure from about 4% less for QTL effect 1.00 (MAF50%) and 9% less for QTL effect 0.50. Regression test was unbiased and bias of GRAMMAR test was presented in figure 5.

Discussion and Conclusion

Regression analysis gave a high false discovery rate with increase of heritability and family structure. Even with random sample in Trotter population, false discovery rate reached 8% with h^2 =0.50 when 5% was expected with unrelated population. The use of families, even of moderate size 30, gave 18% of false discovery rate with this heritability. This high false discovery rate was not synonymous of higher power. In fact, for SNP effect larger than 0.11 (MAF 50%), the power decreased with the increase of heritability and family structure. The increase of power according to family structure was only for small SNP effects, with power lower than 50%. Note that this is the variability between relationships which is responsible for the situation not the mean of relationship between animals (for example the co-existence of sibs and unrelated horses)

GRAMMAR analysis gave lower false discovery rate with related population than without but in a small extend (decrease from 5% without relationship to 2% with large families and high heritability). Even with this test supposed to be corrected for relationships, the power always decreased with the increase of relationships. About 12% power less was expected with a medium heritability of 0.40 and large families (60 progeny by sire). Passed 0.30, the effect of heritability on power was small compared to family structure effect. Bias was very important and mainly depended on heritability rather than on family structure. This is because bias was directly linked to reliability which became independent of the number of progeny with high heritability in presence of own performance. This important bias did not affect the power because the decrease of SNP effect estimate was accompanied by the decrease of residual variance of model (4). But estimates had to be corrected to be expressed in phenotypic scale.

The use of regression analysis gave surely a high number of significant effects with the assurance than most of them are false positives. In that sense, the use of grammar was preferable. The power of grammar was affected in the same magnitude than regression (for sufficiently high SNP effect), so both methods will find QTL with the same probability with related populations. Considering the whole genome, the chance to find QTL with Equine50SNP BeadChip and single SNP analysis is reserved to medium QTL effect, higher than 3/4 phenotypic standard deviation allele effects for equally distributed allele (MAF 50%) and high QTL (1.25 standard deviation) for rare allele (MAF 10%) when considering sample of moderate size (600) but still expensive one.

More investigation will be done about distribution of GRAMMAR test taken into account variability of variance component estimations and departure from Hardy-Weinberg equilibrium. Application to QTDT will also be done to test the behavior of LDLA methods compared to LD methods used here.

References

Hill, EW., Gu, J., Eivers, SS. *et al.* (2010). *PLoS ONE.*, 5(1):e8645.

Aulchenko, YS., de Koning, DJ., and Haley, C. (2007). *Genetics*, 177:577-585.