Statistical modeling for QTL detection

T. Mary-Huard

maryhuar@agroparistech.fr

UMR de Genetique Vegetale, INRA, Universite Paris-Sud, CNRS

UMR AgroParisTech/INRA MIA 518

Outline

QTL identification :

Not the only purpose of the statistical analysis

- Population description (LD, frequencies, kinship...)
- Evaluation of the QTL effect
- Prediction of the genetic value of an individual (genomic selection)

Statistical modeling :

Without being a specialist, one needs to be aware of :

- the biological assumptions implied by the statistical model
- the technical problems that may (will !) be met when applying the model to real data

Caution : technical part inside...

Purpose of statistical modeling

Aim of an experiment : answer to a biological question

Results of an experiment : (numerous, numerical) measurements

Model : mathematical formula that relates the experimental conditions and the observed measurements (response)

(Statistical) modeling : translating a biological question into a mathematical model ($\neq$ PIPELINE !)

Statistical model : mathematical formula involving

- ⋆ the experimental conditions,
- ⋆ the biological response,
- ⋆ the parameters that describe the influence of the conditions on the (mean, theoretical) response,
- ⋆ and a description of the (technical, biological) variability.

Nature of the data

n individuals (lines) described by P markers and a phenotype :

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	...	Y
1	1	0	0	1	1	1	0	...	2.1667951
2	1	1	0	0	0	0	0	...	6.3278859
3	0	0	1	1	0	0	0	...	-0.7408988
4	1	1	0	0	0	1	0	...	3.4949548
5	0	1	1	0	0	0	0	...	2.0677907
6	0	0	0	1	0	0	1	...	-1.1534069
7	0	0	1	1	0	0	0	...	-0.2728285
8	0	1	0	0	0	0	0	...	-3.6509519
9	0	1	1	0	1	0	0	...	-5.1934399
10	0	1	1	0	1	1	0	...	-1.6440094
...									

Y_i : phenotype of individual i

X_{ij} : genotype of individual i at locus j (0 or 1)

Hypotheses

Easy to remove :

SNP = genes, and some of them are QTL. Therefore

⋆ QTL are observed,

⋆ QTL are biallelic.

Uneasy to remove :

No Linkage Disequilibrium !

⋆ Makes simulation easier,

⋆ Makes analysis easier,

But of course unrealistic...

Basics in statistics

Normal distribution

★ Univariate

$$Z \hookrightarrow \mathcal{N}(\mu, \sigma^2) \Rightarrow aZ + b \hookrightarrow \mathcal{N}(a\mu + b, a^2\sigma^2)$$

★ Multivariate

$$Z \hookrightarrow \mathcal{N}(\mu, \Sigma) \Rightarrow AZ + B \hookrightarrow \mathcal{N}(A\mu + B, A\Sigma A^T)$$

Other distributions derived from Normal distribution

★ Khi-2 distribution

$$Z_1, ..., Z_n \hookrightarrow \mathcal{N}(0,1) \text{ i.i.d. } \Rightarrow U = \sum_i Z_i^2 \hookrightarrow \chi^2(n)$$

★ Fisher distribution

$$U_1 \hookrightarrow \chi^2(n_1),\ U_2 \hookrightarrow \chi^2(n_2),\ U_1 \perp\!\!\!\perp U_2 \Rightarrow \frac{U_1/n_1}{U_2/n_2} \hookrightarrow \mathcal{F}(n_1, n_2)$$

Monogenic model analysis in practice

Data :

⋆ 1 phenotypic trait Y

⋆ 1,000 biallelic SNP with $0.1 \leq \text{MAF} \leq 0.9$,

⋆ Among SNPs, 6 QTL with means $\{2, -2, 3, -3, 4, -4\}$,

⋆ 100 individuals.

R program (no package needed)

For each SNP, run :

```
aov <- lm(Y ~ snp.name,data=don)
anova(aov)
```

Monogenic model analysis in practice

★ For a non informative SNP :

```
Analysis of Variance Table
Response: Y
             Df Sum Sq Mean Sq F value Pr(>F)
snp1          1   17.4  17.400  0.9199 0.3399
Residuals    98 1853.7  18.915
```

★ For an informative SNP (ie a QTL) :

```
Analysis of Variance Table
Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
snp161     1  219.12 219.118  12.999 0.0004923 ***
Residuals 98 1651.98  16.857
```

Overall, 51 SNP with significative pvalues...

Testing simultaneously K hypotheses

With one hypothesis :

		Decision	
		H_0 accepted	H_0 rejected
Truth	H_0 true	TN	FP
	H_0 false	FN	TP

With K hypotheses :

		Decision			
		H_0 accepted		H_0 rejected	
K_0	H_0 true	TN	true negatives	FP	false positives
K_1	H_0 false	FN	false negatives	TP	true positives
K hypotheses		N negatives		P positives	

The multiple testing problem

Assume

⋆ K_0 hypotheses H_0 are true

⋆ all tests are performed at level α

Define

$$
\begin{array}{lll}
Z_1 = 1 & \text{if } p_1 \leq \alpha, \text{ else } Z_1 = 0 & \Rightarrow \quad Z_1 \hookrightarrow \mathscr{B}(\alpha), \\
Z_2 = 1 & \text{if } p_2 \leq \alpha, \text{ else } Z_2 = 0 & \Rightarrow \quad Z_2 \hookrightarrow \mathscr{B}(\alpha), \\
\ldots & & \\
Z_{K_0} = 1 & \text{if } p_{K_0} \leq \alpha, \text{ else } Z_{K_0} = 0 & \Rightarrow \quad Z_{K_0} \hookrightarrow \mathscr{B}(\alpha),
\end{array}
$$

$$
FP = \sum_{k=1}^{K_0} Z_k \qquad \Rightarrow \quad \sum_{k=1}^{K_0} Z_k \hookrightarrow \mathscr{B}(K_0, \alpha).
$$

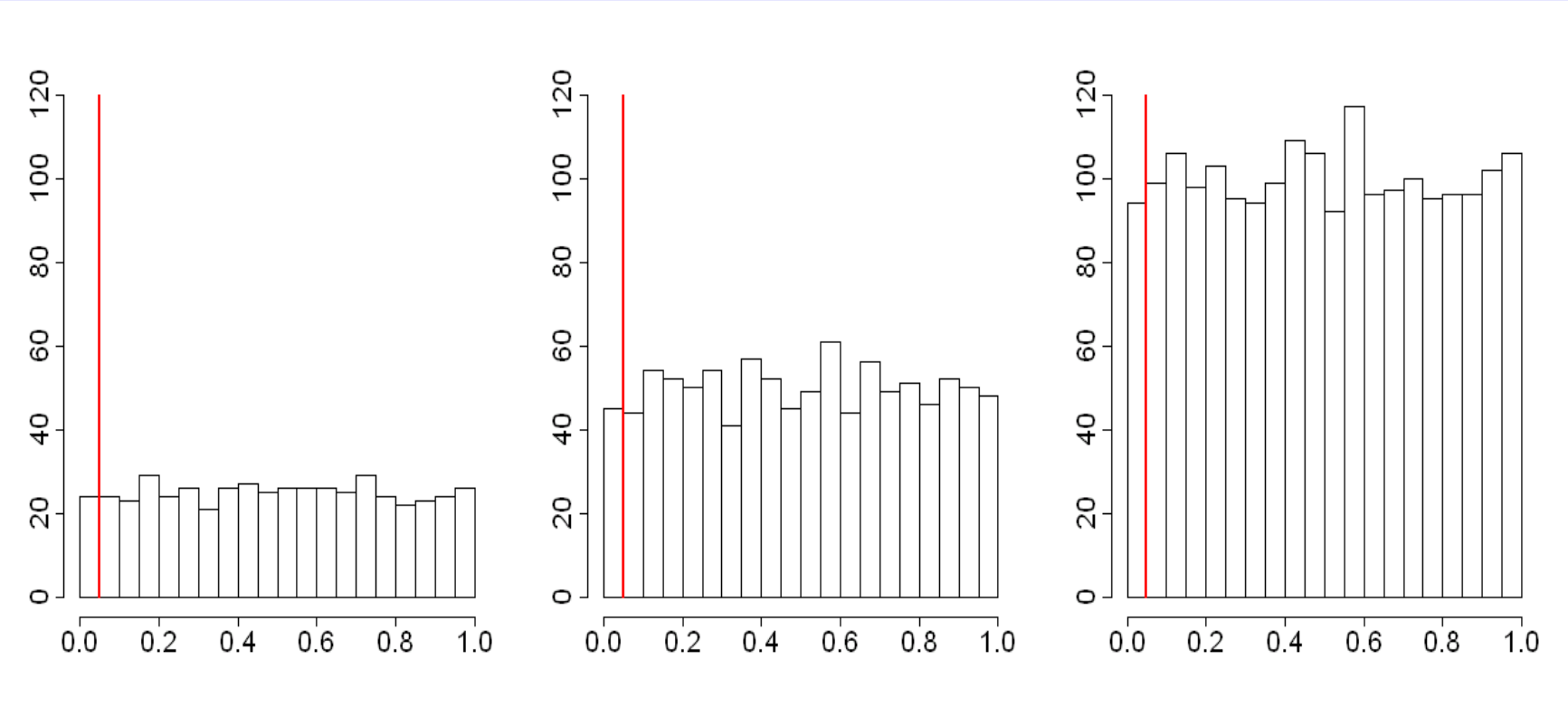
In particular, one has

$$
\mathbb{E}(FP) = K_0 \alpha
$$

⇒ **The expected number of false positives increases linearly with K_0...**

The multiple testing problem

Illustration ($\alpha = 0.05$)



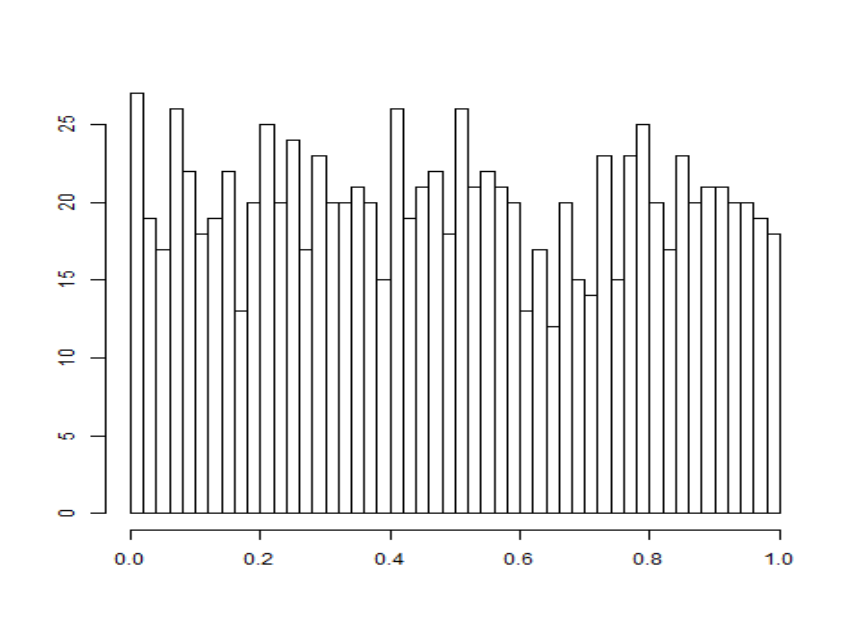
$K_0 = 500$
$FP = 25$

$K_0 = 1000$
$FP = 45$

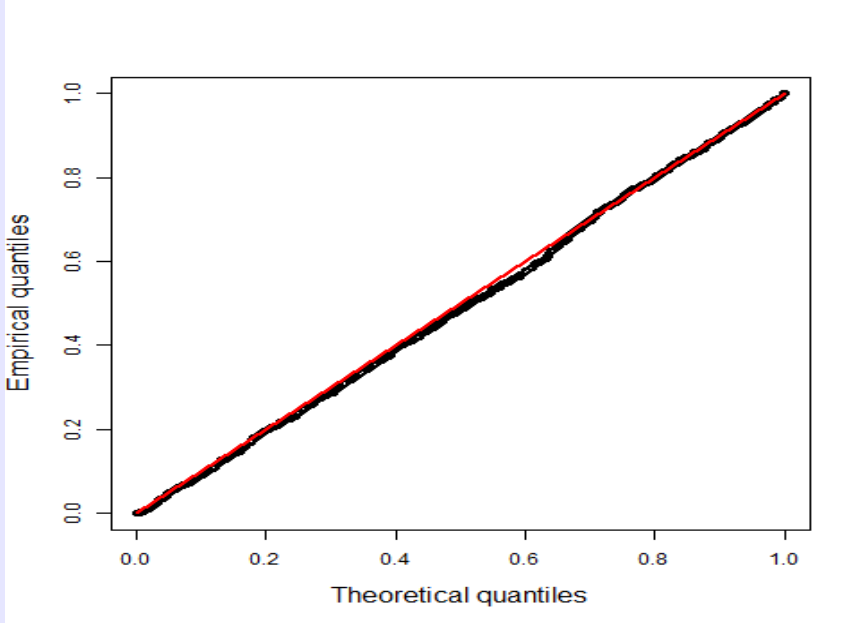
$K_0 = 2000$
$FP = 94$

Monogenic model analysis in practice

Diagnostic on pvalue distribution :



Histogram



QQ-plot

Bonferroni correction : 2

BH correction : 2

Oligogenic model analysis in practice

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0867     0.4565   0.190     0.85
snp109        1.7964     0.3017   5.955 4.58e-08 ***
snp377       -1.8016     0.2692  -6.693 1.62e-09 ***
snp332        4.7054     0.2711  17.354  < 2e-16 ***
snp124       -4.4076     0.2844 -15.495  < 2e-16 ***
snp120        3.0409     0.2725  11.160  < 2e-16 ***
snp161       -3.1319     0.2751 -11.385  < 2e-16 ***

Residual standard error: 1.256 on 93 degrees of freedom
Multiple R-squared:  0.9216,  Adjusted R-squared:  0.9165
F-statistic: 182.1 on 6 and 93 DF,  p-value: < 2.2e-16
```

GWAS model analysis in practice

Data :

⋆ 1 phenotypic trait Y

⋆ 1,000 biallelic SNP with $0.1 \leq \text{MAF} \leq 0.9$,

⋆ Among SNPs, 6 QTL with means $\{2, -2, 3, -3, 4, -4\}$,

⋆ 100 individuals : 20 families of 5 relatives,

⋆ Mean level of kinship between relatives : small or high

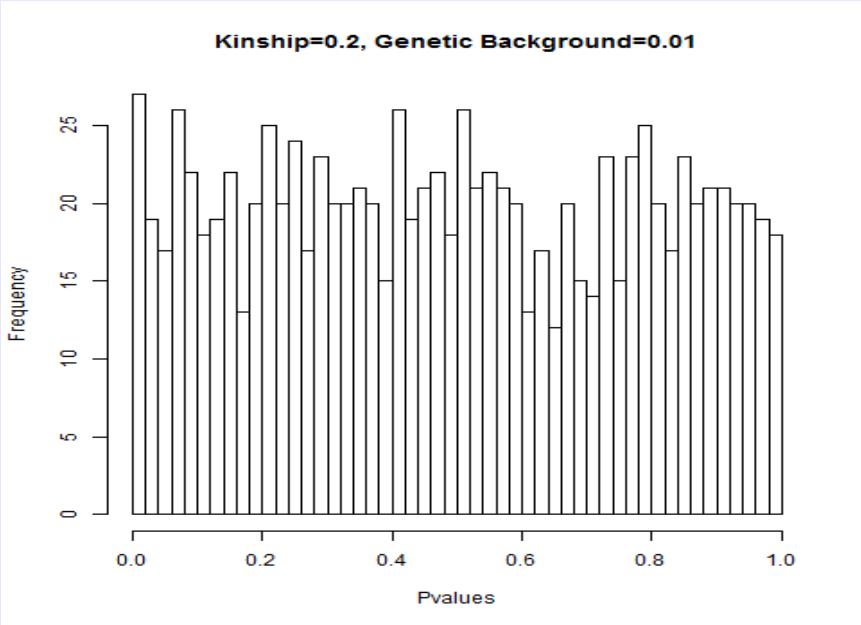
⋆ Genetic background : all genes (except QTLs) have small/moderate effects.

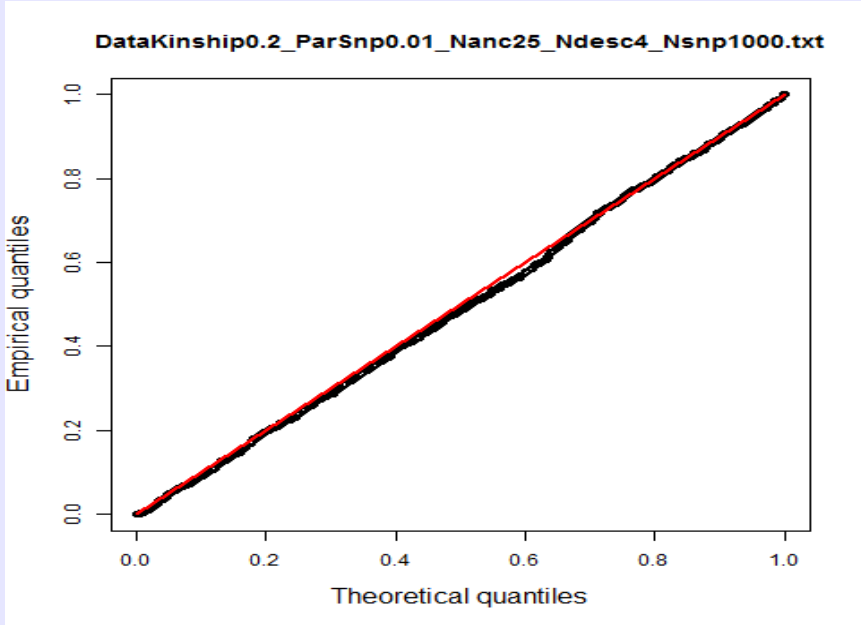
R program (with Asreml)

For each SNP, run :

```
gwas <- asreml(Y ~ snp, random = ~ giv(Ind,var=T),
               ginverse=list(Ind=InvKinship), data=don)
summary(res)$varcomp$component
anova(res)
```

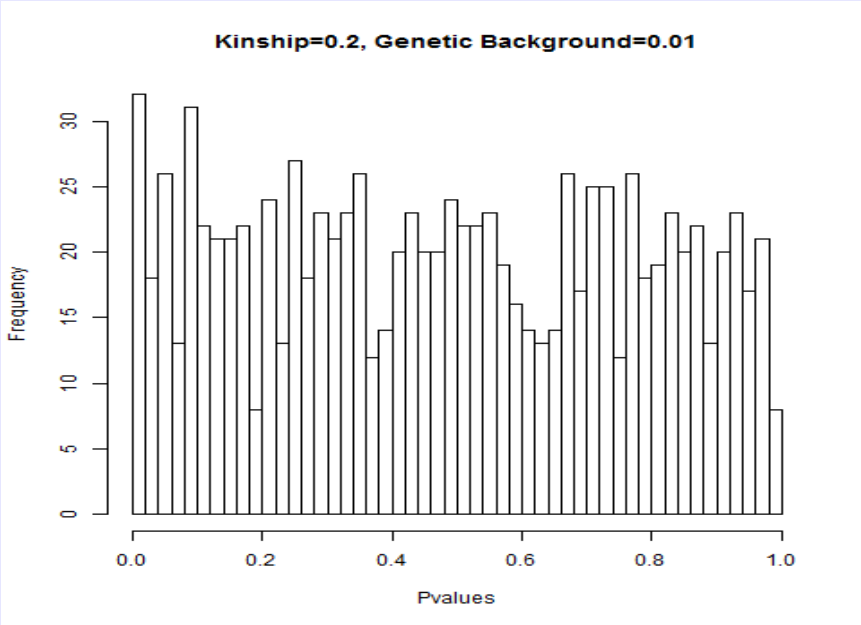
Remember monogenic model analysis ?

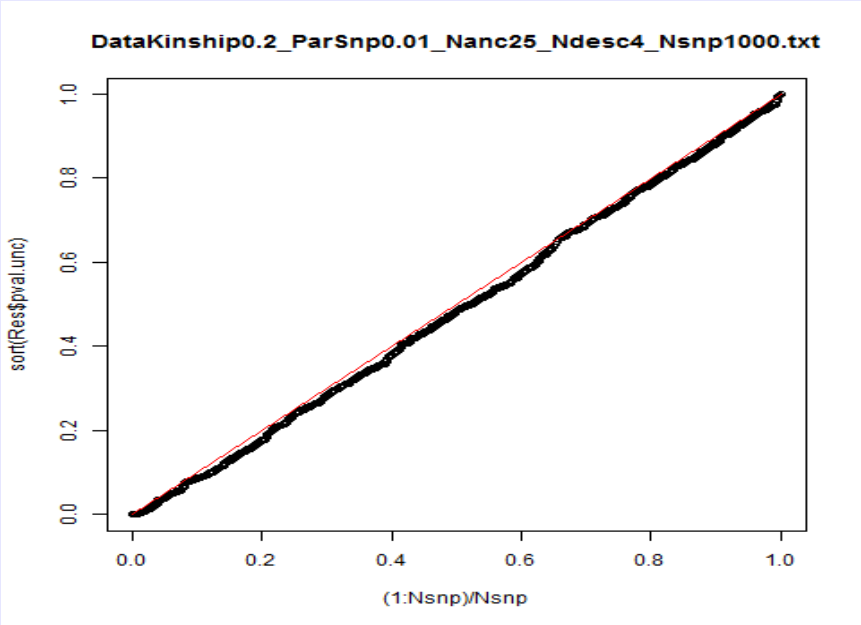




	Declared QTL	True QTL	Proportion	Missing
No correction	51	6	0.88	0
Bonferroni correction	2	2	0	4
BH correction	2	2	0	4

Same data, GWAS model analysis

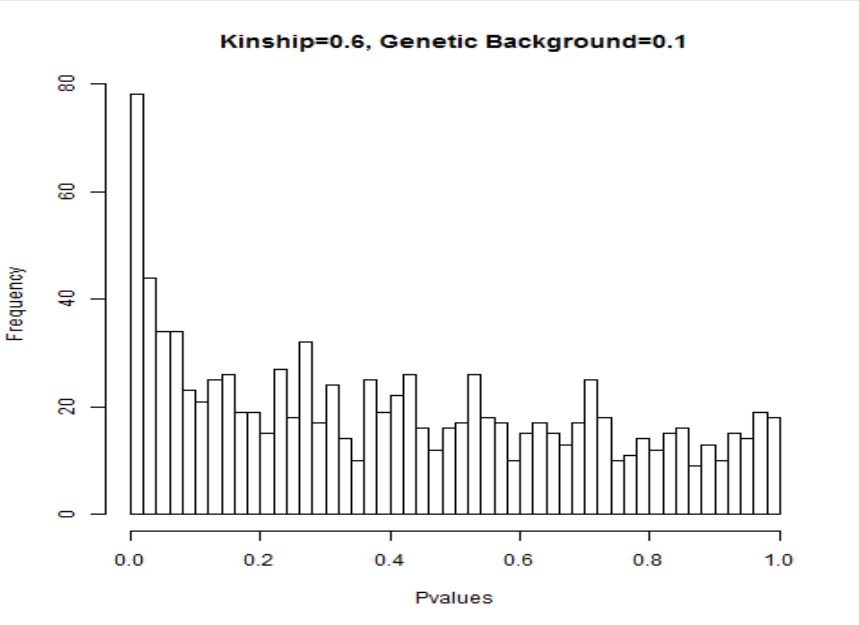


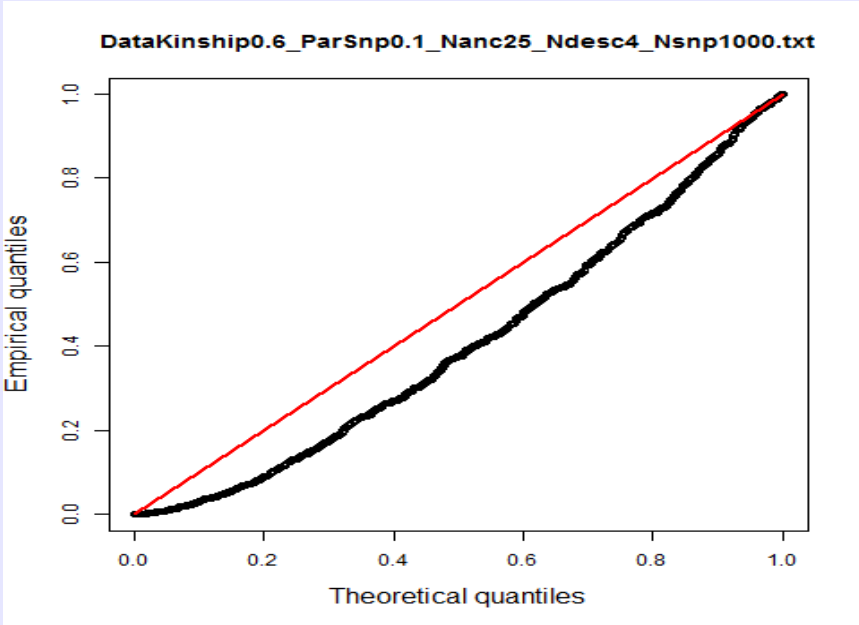


	Declared QTL	True QTL	Proportion	Missing
No correction	62	6	0.90	0
Bonferroni correction	2	2	0	4
BH correction	2	2	0	4

High level of kinship and genetic background

Monogenic data analysis

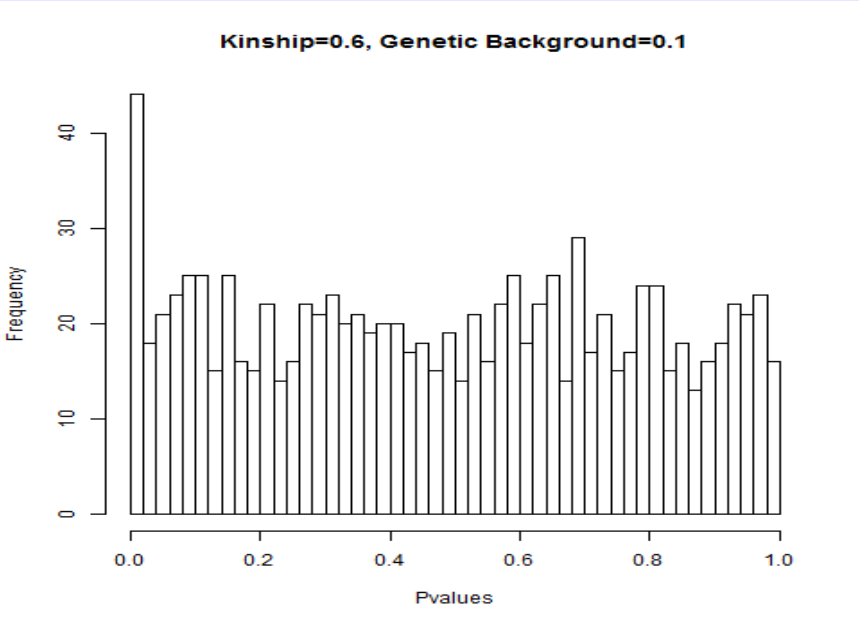


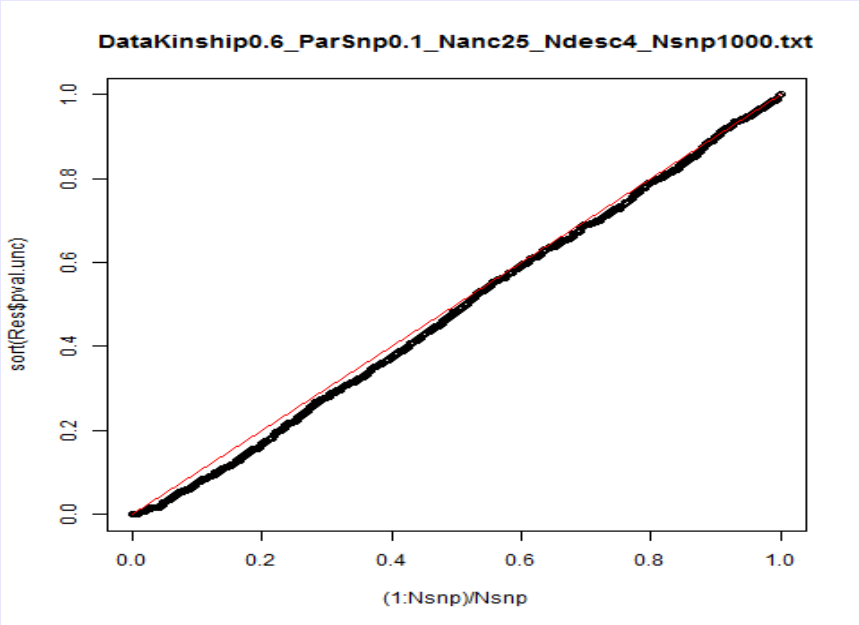


	Declared QTL	True QTL	Proportion	Missing
No correction	141	4	0.88	0
Bonferroni correction	5	3	0.4	3
BH correction	10	4	0.6	2

High level of kinship and genetic background

GWAS data analysis





	Declared QTL	True QTL	Proportion	Missing
No correction	69	5	0.88	0
Bonferroni correction	3	3	0	3
BH correction	3	3	0	3