ENHANCED PRECISION FOR QTL-MARKER PARAMETER ESTIMATION WITH SELECTIVE DNA POOLING

C. Carleos[1], N. Corral[1], J.A. Baro[2] and J. Cañon[3]

[1] Depto. Estadística, Escuela de Marina, 33203 Spain
[2] Depto. Ciencias Agroforestales, ETSIA, 34004 Spain
[3] Depto. Produccion Animal, Facultad de Veterinaria, 28040 Spain

INTRODUCTION

QTL-marker association studies often involve genotyping of large samples. Selective DNA pooling (Darvasi and Soller 1994) can reduce dramatically genotyping costs. The technique consists of selecting animals in groups according to their phenotypes, and making a pool (one per group) with tissue samples from all animals of the group and, finally, pools are typed. Analysis of allelic band intensity provide estimates of allele frequencies.

The main drawback of selective DNA pooling in comparison with individual genotyping, comes from technical error. It originates either from unequal contributions of individuals to the pool, or from artefacts concerning band analysis, such as measurement errors, differential amplification, etc. These may be termed laboratorial or "physical" sources of imprecision.

An additional source of imprecision is the intrinsic nature of data provided by DNA pooling. The genotypic information is fully separated of phenotypes. The assignment of genotypes to phenotypes is not possible, unlike individual genotyping. Further, traditional approaches do not extract from the sample all the information about QTL-marker association: they incorporate phenotypic and genotypic data in an unintegrated manner. In this work, we try to make use of all the information provided by selective DNA pooling, in order to minimize the effect of this source of imprecision.

MATERIAL AND METHODS

An oversimplified framework will be considered. Phenotypic records are observed for N individuals. The genotype of individual i is x_i, and its phenotype is y_i. Only two possible genotypes are present, denoted as 0 and 1. Phenotype is a continuous trait with probability density ϕ, depending on the parameter θ; this one takes two possible values according to the genotype. A selection threshold l is set such that all animals with trait values below l will be genotyped, either individually or in a pool.

From now onwards, we will use "individual method" and "pooled method" to refer to traditional methods of analysis handling individual selective genotyping and selective DNA pooling; and will use "individual data" and "pooling data" to refer to sample data obtained from either strategy. A proposed new method will be referred to as P2I (pooling to individual).

Let $L:=\{i \mid y_i<l\}$ represent the selected lower tail. Individual selective genotyping implies knowledge of x_i, $i \in L$. Selective pooling only provides an estimate of $n:=\Sigma\{x_i \mid i \in L\}$; that estimate is assumed exact along this work.

The likelihood of an individual genotyping is :

$$\Lambda((x_i)_{i=1}^{L},(y_i)_{i=1}^{N} \mid \theta) \propto \prod_{i=1}^{L} \phi(y_i \mid x_i,\theta) \prod_{i=L+1}^{N} \phi(y_i \mid \theta)$$

The complete, unmanageable likelihood for pooling data follows :

$$\Lambda(n,(y_i)_{i=1}^{N}) \propto \sum_{\sigma \in S_L} \Lambda((x_{\sigma(i)})_{i=1}^{L},(y_i)_{i=1}^{N} \mid \theta) \qquad \text{(Eq. 1)}$$

where the sum is over all combinations of individual genotypes compatible with the pooling data. The traditional approach replaces that formula by

$$\Lambda(n,(y_i)_{i=1}^{N}) \propto \Phi(l \mid X=0,\theta)^{L-n}\ \Phi(l \mid X=1,\theta)^{n}\ \left[1-\Phi(l \mid \theta)\right]^{N-L} \prod_{i=1}^{N} \phi(y_i \mid \theta)$$

Note that any information about individual genotype/phenotype relationship is discarded.

Our proposal is an iterative method to approximate the sum in Eq. 1 by repeating two steps (Tanner 1996) :

a) Parameter step: an estimate is obtained from the likelihood "augmented" in step b), e.g., the likelihood is maximized with respect to the parameters of interest, θ.

b) Imputation step: Monte Carlo draws of "latent" genotypes according to their conditional distribution on the pooling data and the parameter values of step a), p(X | θ, Y, n). Here, X represents individual genotypes, Y are phenotypes, n as above.

Distribution p(X | θ, Y, n) in step b) is

$$p(X \mid \theta, Y)) = \frac{\prod_i \phi(y_i \mid x_i,\theta)}{\Sigma\left\{\prod_i \phi(y_i \mid x_i',\theta) \mid \#\left\{x_i'=1\right\}=n\right\}}$$

so latent genotypes must be drawn with probability proportional to $\Pi\{\phi_i|x_i=1\}$, with $\phi_i:=\phi(y_i|x_i=1) / \phi(y_i|x_i=0)$.

A first attempt used a typical rejection/acceptance algorithm, but it proved to be impractical. The cause is that any series of bounds of the probability of latent genotype sample, turned out to be much higher than the probability of most possible samples. Thus the algorithm spends a lot of time to draw each sample. An alternative way is sampling without replacement n integers from 1 to L. Sampling must be with unequal probabilities in order to obtain a sample s such that Pr(s) is proportional to $\Pi\{\phi_i|x_i=1\}$. The following scheme does so:

1- Assign to $1 \le i \le L$ a probability $z_i:=c\ \phi_i / (1 + n\ c\ \phi_i)$, where c is a normalizing constant, so that z_i sum to one.

2- Apply Sampford's method (Cochran 1973) to obtain a sample *s*.
2.1- Draw the first unit *i* with probability z_i.
2.2- Select another unit $j \in \{1, ..., L\} \setminus \{i\}$ with probability proportional to $z_j / (1 - n\, z_j)$.
2.3- If *j* is equal to any of the previously drawn units, all the sample (except *i*) is rejected.
2.4- If the sample size is less than *n*, go to 2.2.
Note that the method samples with replacement, in order to obtain a sample without replacement.

3- With that method, the sample *s* is drawn with probability

$$\Pr[s] \propto \prod_{i \in s} z_i \frac{1 - \sum_{i \in s} z_i}{\prod_{i \in s} (1 - n\, z_i)} = (1 - \sum z_i) \prod \frac{z_i}{1 - n\, z_i}$$

so the last term must be corrected.

4- Let

$$b := \frac{1}{1 - \sum \{ z_{(i)} \mid i = L - n + 1, ..., L \}}$$

where $z_{(i)}$ is the *i*-th element of *z* in ascending order.

5-Let *u* be a U(0, *b*) random number. If $u > 1/(1 - \sum \{z_i | i \in s\})$ then reject *s* and go to 2.

RESULTS AND DISCUSSION

The three methods were compared for several combinations of parameters. Simulations were carried out by assuming ϕ being a normal distribution, $N(\mu - \alpha/2, \sigma)$ for genotype 0 and $N(\mu + \alpha/2, \sigma)$ for genotype 1. Where not stated otherwise, parameters take the following values:

-*N*=1000, sample size
-μ=150, overall mean
-σ=10, residual standard deviation
-α=σ, QTL effect
-*l*=μ-σ, selection threshold

Table 1 shows standard errors for estimators of α; departures from the previous values are indicated as well. 10000 realizations were performed for each combination of parameters. Estimates turned out to be unbiased.

Table 1. Standard errors of QTL effect estimators with individual, pooling and P2I methods

Individual	Pooling	P2I	Departure from initial values
0.829	0.876	0.852	none
1.838	1.952	1.872	N=200
1.189	1.251	1.223	N=500
1.337	1.363	1.349	N=500, $\alpha=\sigma/2$
0.937	0.986	0.959	$\alpha=\sigma/2$
0.967	1.027	0.991	$\alpha=\sigma/4$

It can be seen that performance of P2I method is intermediate between individual and pooling. A certain amount of information is recovered from pooling data with respect to the traditional approach.
The method presented here is not restricted to use maximum likelihood. The parameter step may involve any method developed to deal with individual data. Obtained values are adequately combined to give the result corresponding to pooling data.

CONCLUSION

An iterative procedure is proposed to extract the maximum of information from a selective DNA pooling sample. It achieves a moderate performance enhancement in comparison with traditional pooled genotyping analysis. On the other hand, it allows a direct employment of strategies based on individual genotyping, on pooled genotyping data.

REFERENCES

Cochran, W.G. (1977) “Sampling techniques” John Wiley and Sons, New York, U.S.A.
Darvasi, A. and Soller, M. (1994) *Genetics* **138** : 1365-1373.
Tanner, M.A. (1996) “Tools for statistical inference” Springer-Verlag, New York, U.S.A.