

Nonparametric Iterated Procedure for Testing Genetic Differentiation

*Procedura Non Parametrica Iterata
per Saggiare la Differenziazione Genetica*

Livio Finos, Fortunato Pesarin, Luigi Salmaso, Aldo Solari
Dipartimento di Scienze Statistiche, Università di Padova. solari@stat.unipd.it

Riassunto: In questo lavoro viene presentata una soluzione di permutazione basata sulla teoria della combinazione non parametrica di test dipendenti proposti da Pesarin (2001) al problema della differenziazione genetica. La capacità di rilevare plausibili eterogeneità genetiche combinando l'informazione di più loci differisce a seconda del metodo impiegato, cosicché nessun test statistico domina uniformemente gli altri. La procedura proposta risulta robusta e flessibile rispetto alle possibili configurazioni dell'ipotesi alternativa globale.

Keywords: genetic differentiation, permutation test, Hardy-Weinberg equilibrium, linkage equilibrium, nonparametric combination methodology, iterated procedure.

1. Introduction

Statistical tests for genetic differentiation are increasingly common in conservation and evolutionary biology. It is tested whether or not observed differences in allele frequencies between samples are large to conclude that the samples do not represent the same biological population or gene pool (Ryman and Jorde, 2001). In order to answer this question, two samples of individuals, separated in space or time, are genotyped at multiple loci, and each sample is described in terms of its size and allele frequencies. The general statistical approach most frequently used is to firstly determine a contingency test for allele frequency homogeneity for each locus separately, and then to evaluate the simultaneous, or joint, information from all examined loci. The question of how to combine information from multiple loci is increasingly significant because the current trend in many fields of evolutionary biology appears to be to score a steadily growing number of loci in a fairly restricted number of individuals, frequently in the range, say 5-30 (Ryman and Jorde, 2001).

2. Standard methods and iterated procedure

Ryman and Jorde (2001) compare the power of two conceptually different strategies used to test the joint null hypothesis of no difference at any locus - the first proposed method is based on the summation of chi-square statistics over loci (summation method) and the second is based on the Bonferroni adjusted Chi-square test.

Into permutation methods we introduce a robust test that is relatively powerful for all possible alternatives; this solution is obtained through the Nonparametric Combination

methodology (NPC, Pesarin, 2001). The responses of the problem are binary, so that, $X_{hji} = 0$ or 1 , $i = 1, \dots, 2 \cdot n_j$, $j = 1, 2$, $h = 1, \dots, k$, where h is the locus, j is the sample and i is the gene. Note that individuals are diploid, thus each individual has two alleles at any locus, so there will be twice as many alleles as statistical units. It is also assumed that the Hardy-Weinberg equilibrium holds, thus alleles within individuals are independent, and the analysis that treats alleles rather than genotypes as individual entities is appropriate (Sasieni, 1997).

The hypotheses under testing are $H_0 = \left\{ \bigcap_{h=1}^k [X_{h1} \stackrel{d}{=} X_{h2}] \right\} = \left\{ \bigcap_{h=1}^k H_{0h} \right\}$ against the alternative $H_1 = \left\{ \bigcup_{h=1}^k [X_{h1} < \stackrel{d}{\neq} X_{h2}] \right\} = \left\{ \bigcup_{h=1}^k H_{1h} \right\}$; where the notation $< \stackrel{d}{\neq} >$ indicates either stochastic dominance or inequality relations. Appropriate permutation partial tests are:

$$T_h^* = \left(\frac{\sum_{i=1}^{2 \cdot n_2} X_{h2i}^*}{2 \cdot n_2} \right) - \left(\frac{\sum_{i=1}^{2 \cdot n_1} X_{h1i}^*}{2 \cdot n_1} \right) \quad (1)$$

for the two-sided sub-alternative or, for example, $\bar{T}_h^* = \sum_{i=1}^{2 \cdot n_2} X_{h2i}^*$ for the one-sided sub-alternative $H_{1h} : X_{h1} < \stackrel{d}{\neq} X_{h2}$.

In the framework of permutation methods it is possible to think about any kind of combining function in two stages - first the combination of k partial tests into a global test, and then the combination of two, or even more, global tests all from the first stage. The second stage may become necessary because different combining functions, i.e. different rejection regions, may of course give slightly different overall p -values, although, due to their consistency, they are asymptotically equivalent in the alternative (Pesarin, 2001); however, in order to reduce their influence, we may *iterate* the combination procedure. In fact, following Pesarin's practical guidelines (2001), we should choose Tippet's, Fisher's and Liptak's combining function when we expect one or few, several, and almost all loci respectively to be under the alternative. Due to the lack of prior information, our preference is for a robust combination: we obtain a powerful global test both when there are differences in many or all loci and when there are differences regarding only one or a few loci. We briefly present the two-stage test where we consider Tippet's second stage combining function:

$$\underbrace{\begin{matrix} T_T^{**} = \max_{1 \leq i \leq k} (1 - \lambda_i) \\ T_F^{**} = -2 \sum_{i=1}^k \log(\lambda_i) \end{matrix}}_{\text{first stage}} \xrightarrow{\text{iterated procedure}} \underbrace{T_T^{***} = \max(1 - \lambda_T'', 1 - \lambda_F'')}_{\text{second stage}} \quad (2)$$

In the following section we present a simulation study that takes three first stage global tests into account: the further combining function considered is Liptak's $T_L^{**} = \sum_{i=1}^k \Phi^{-1}(1 - \lambda_i)$ that behaves better than Fisher's combining function if (almost) all the k one-sided alternatives are jointly true with the same standardized non-centrality parameter, a condition that is rarely satisfied in many realistic situations.

3. Comparative simulation study

To compare the different methods two simulation studies were performed: the first shows the behaviour of the considered combining functions, the second is a comparative study with the results from Ryman and Jorde (2001).

Random sampling of $2 \cdot n$ genes (n diploid individuals) from a population with known allele frequencies has been simulated; the size and power of the tests are estimated using a 5% significance α -level. For each configuration, we performed 2000 simulations and 1000 Conditional Monte Carlo iterations (CMC, Pesarin, 2001).

3.1 Power behaviour of combining functions

We present results for the case of 12 di-allelic loci and $n = 50$ diploid individuals (100 genes) for each sample; the allele frequency for each locus is $Q_1 = 0.10$ in the 1st sample, and $Q_{h2} = 0.10$, $h = 1, \dots, 12$, if the locus h is under H_0 , otherwise $Q_{h2} = Q_2$ (see Table 1). For simplicity and without loss of generality, we consider all one-sided sub-alternatives as $H_{1h} : X_{h1} <^d X_{h2}$, $h = 1, \dots, 12$ and the global alternative $H_1 = \left\{ \bigcup_{h=1}^{12} [X_{h1} <^d X_{h2}] \right\}$.

The test statistic considered for the simulation study is $\sum_{h=1}^{12} X_h^2$, i.e. the sum of a set of conventional Chi-square 2×2 contingency statistics ($X_h^2 \sim \chi_1^2$), that follows a χ_k^2 distribution with d.f. equal to the sum of the d.f. of the contributing independent variables. Note that if there is dependence among loci (linkage disequilibrium) the summation method is invalid ($k \cdot \chi_{\alpha,1}^2 \geq \sum_{h=1}^k X_h^2 \geq \chi_{\alpha,k}^2$; left bound is reached in the case of complete dependence, right bound in the case of independence) whereas the proposed test is valid because the combination was done nonparametrically with regard to the underlying dependence relations. The results are listed in Table 1, and agree with our expectations. From a series of Monte Carlo experiments we found that the Liptak combining function does not behave better than the Fisher combining function for the two-sided alternative.

Table 1: Power estimates

loci(•) sub H_1	Q_2	$\bar{T}_T^*$	$\bar{T}_F^*$	$\bar{T}_L^*$	$\bar{T}_T^{**}$	$\sum X_h^2$
•○○○○○○○○○○○○	0.30	.8440	.4027	.2140	.5827	.6487
●○○○○○○○○○○○○	0.26	.8627	.8167	.4447	.8400	.8000
●●○○○○○○○○○○	0.22	.7380	.8273	.5713	.8233	.7567
●●●○○○○○○○○	0.20	.6647	.8880	.7200	.8720	.7487
●●●●○○○○○○○	0.20	.7567	.9573	.8720	.9487	.8660
●●●●●○○○○○○	0.18	.6107	.9293	.8720	.9227	.7573
●●●●●●○○○○○	0.18	.6747	.9753	.9413	.9640	.8340
●●●●●●●○○○○	0.18	.7127	.9920	.9807	.9887	.8993
●●●●●●●●○○○	0.16	.5067	.9380	.9380	.9320	.6993
●●●●●●●●●○○	0.16	.5367	.9673	.9740	.9660	.7487
●●●●●●●●●●○	0.16	.5807	.9833	.9880	.9840	.8027
●●●●●●●●●●●	0.16	.5907	.9947	.9987	.9940	.8420

3.2 Comparative study

We replicate the study of Ryman and Jorde (2001) that considers the case of k di-allelic loci ($k = 10, 20, 30$) and $n = 20$ diploid individuals (40 genes) per sample; the allele frequency for each locus is $Q_1 = 0.10$ in the 1st sample, and $Q_2 = 0.15$ in the 2nd sample. We consider the two-sided global alternative $H_1 = \left\{ \bigcup_{h=1}^{l^2} [X_{h1} \neq^d X_{h2}] \right\}$ for comparison with the summation method. Simulation results are shown in Figure 1. We also perform a simulation (Figure 2) with different frequencies within a population (in this simulation in the 2nd sample the allele frequency is 0.35 for 10% of loci) because Ryman and Jorde (2001) wrote that «in the real world it is not very likely to encounter a situation where all loci examined in a population segregate at exactly the same frequency». The results show the robust behaviour of the iterated procedure.

Figure 1: 100% Active loci

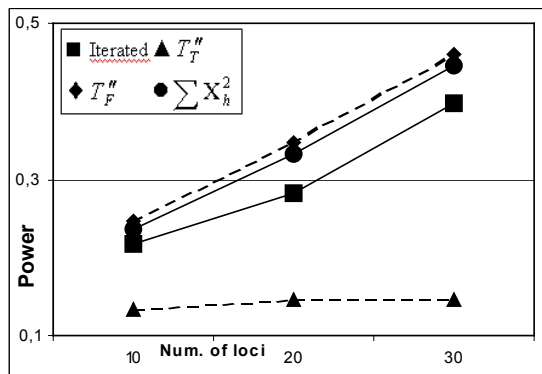
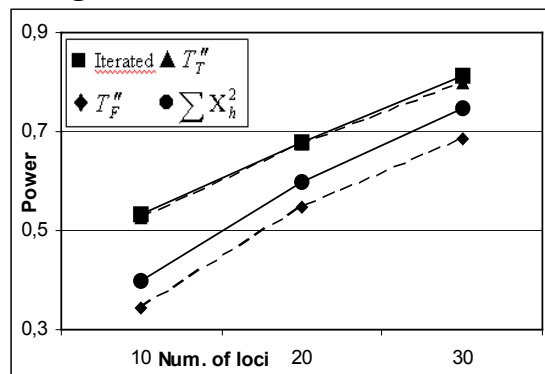


Figure 2: 10% Active loci



5. Conclusions

In this paper we presented a robust solution in testing genetic differentiation. The conclusions that emerge from the above simulations show that the proposed iterated procedure exhibits a good overall performance, also in the case of linkage disequilibrium, whereas the summation method becomes invalid in the case of dependence among loci. We would emphasize the versatility of NPC, particularly if *a priori* information is available, allowing restricted alternative analysis, stratified, weighted and closed testing procedures.

References

- Pesarin F. (2001) *Multivariate permutation tests with applications in biostatistics*. Wiley, Chichester.
- Ryman N., Jorde, P.E. (2001) Statistical power when testing for genetic differentiation. *Molecular Ecology*, 10, 2361-2373.
- Sasieni P.D. (1997) From genotypes to genes: Doubling the sample size. *Biometrics*, 53, 1253-1261.