Adaptive permutations

We recommend the reader to refer to Che R. *et al.*, 2014.

Permutation-testing is a non-parametric (distribution-free) strategy to assess statistical significance. Although powerful, this approach is computationally inefficient and its execution-time tends to be prohibitive when the approach is applied to large datasets, as it is the case in GWAS applications.

The standard form of permutation testing is based on a randomization procedure. The response variable (*i.e.* the phenotype) is shuffled *n* times, which leads to the creation of *n* permuted datasets. For each permuted dataset, the test statistic (a G-test for instance) is computed and recorded. All statistics then generate a distribution representing the probability distribution of observing the genotypes under the null hypothesis of no association between genotypes and phenotype. To perform the statistical test and assess the significance of the association, the observed statistic has to be compared to the generated distribution. If it is the R^{th} largest statistic value among the distribution, then the p-value is equal to *R/n*.

Performing correction for multiple testing requires many permutations and then becomes computationally prohibitive on large GWAS datasets, *a fortiori* in epistasis analysis where many genetic marker combinations have to be tested. That is why SMMB uses a computationally efficient adaptive permutation approach to perform multiple test corrections. Adaptive permutations are faster to compute than standard permutations because they reduce greatly the total number of permutations needed to estimate the p-value of multiple permutation tests.

The pool of tests included in the procedure may be reduced by discarding tests whose estimated p-values are larger than the currently observed p-value. As soon as *r* permuted tests result in a test statistic smaller than the observed test statistic, all permutations that would have remained for the current test are discarded. The relation between *r* and α, the global type I error threshold specified by the user, is detailed in (Che R. *et al.*, 2014). An estimated p-value is then computed for the current test as $p_{estimated} = r/N$, where *N* is a random variable computed by a censored negative binomial distribution $N\ truncNB(r, \alpha, n)$. Otherwise (if no such *r* permuted tests can be reached), the p-value is estimated in the same way than in a standard permutation testing approach.

Bibliographical references

Che R. *et al.* (2014) An adaptive permutation approach for genome-wide association study: evaluation and recommendations for use. *BioData Min.* 7: 9. doi: 10.1186/1756-0381-7-9.