Composite hypothesis testing procedure for the integration of multiple GWAS summary statistics

Annaïg De Walsche*†[1]

[1]Institut Pasteur – Institut Pasteur de Paris – France

Abstract

Data integration often involves analysing results from different experiments to identify complex patterns. In this context, we consider a scenario where we have a collection of elements $i=1,. . . ,n$ (genes, for example) for which the hypotheses : "element i has no effect in condition q" have been tested for Q conditions. Each observation i, therefore, consists of a vector of Q critical probabilities. The analysis aims to identify the elements that have an effect in all the conditions or a predefined subset of those conditions. The critical probabilities must then be combined flexibly to explore complex hypotheses, called composite hypotheses, while controlling the false positive rate.
We propose a composite hypothesis testing procedure based on a model where the Q-uplet of p-value associated with each gene/marker is distributed as a multivariate mixture where each of the 2^Q components corresponds to a specific combination of H_0 and H_1 states. Our method explicitly accounts for the dependence structure across p-value series through a copula function. The inference of this 2^Q component mixture model is performed efficiently, allowing its application to cases where the number of markers is 10^5–10^6 and where $Q=20$. The inference procedure consists of two independent steps: first, fitting a non-parametric mixture model to the marginal distribution of each Q series of p-values, then estimating the proportions of the mixture model components and the copula parameters using an EM algorithm. Step (E) is optimised to reduce the memory burden of the procedure from $O(n \times 2^Q)$ to $O(n + 2^Q)$. *Applications on simulated data have been carried out, with conclusive resu*

Keywords: Next generation hypothesis testing

*Speaker

†Corresponding author: annaig.de-walsche@pasteur.fr