Exact distribution of genetic risk score for estimating personal risk in complex diseases

Juan R González[1], *Isaac Subirana*[2], *Gavin Lucas*[2], *Carla Lluís-Ganella*[2], *Roberto Elosua*[2]

[1]jrgonzalez@creal.cat, Bioinformatics Research Group in Epidemiology (BRGE), Center for Research in Environmental Epidemiology (CREAL), Barcelona, Spain

[2]isubirana / glucas / clluis / relosua @imim.es, Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain

In recent years, genome wide association studies have identified hundreds of variants, mainly SNPs (Single Nucleotide Polymorphism), associated with a range of complex phenotypes including height, body mass index, or important diseases, such as cancer and heart disease, among others. The most striking general observation of these studies is the fact that the effects of the variants identified are generally weak and have a very low predictive power. In order to improve such performance, several authors have explored the effect of accumulating different associated risk variants on disease risk. Therefore, a genetic score is built by summing the total number of risk alleles (0, 1 or 2) that an individual is carrying for each of the SNP associated with the disease. While these empirical studies highlight interesting relationships between genetics and complex phenotypes, the utility of genetic scores in public health is yet to be established, and, in part, this is due to our reliance on empirical data for exploring additive genetic effects in the population. Knowing the distribution of a genetic score can be useful to make risk prediction estimates or discriminate individuals having different disease susceptibility. Nonetheless, the exact distribution of a genetic score at the population level is unknown.

With the aim of contributing to the set of tools available for exploring the population dynamics and clinical potential of genetic scores, we introduce a new methodology based on the sum of binomial random variables methods for computing the exact population distribution of a genetic score based only on the frequencies of its component risk alleles. The performance of our approach is checked under the assumption of independence between alleles at each SNP (e.g. Hardy Weinberg Equilibrium), and between SNPs included in the genetic score (e.g. Linkage Disequilibrium (LD)). We also demonstrate the usefulness of our method by using a real example in which a set of SNP related to obesity are used. By using this example we show how individuals can be stratified in different risk groups depending on high risk quantiles from exact genetic score distribution. We also provide a bioinformatic tool to get risk profiles having the group of SNPs belonging to the genetic risk score from which the risk alleles can be obtained from existing databases such as 1000 Genome projects to get accurate estimates in the population of interest.

Keywords: Statistical genetics, SNPs, genetic score, exact distribution, sum of binomials.