A statistical method identifies significant biomolecular differences in population genetics

M. G. Borges[1,2]
[1]Deptament of Medical Genetics, FCM, UNICAMP, [2]Brazilian Institute of Neuroscience and Neurotechnology.

Introduction: With the introduction of next-generation-sequencing it comes feasible to identify millions of variants in a single experiment. In this context, an important computational challenge is to correctly annotate these variants for downstream analysis in order to prioritize and filter variants with biological and clinical significance. Currently, filtering parameters used are variant allele frequencies, genomic location or previous disease associations. Cut-off points have been well established for these parameters in the context of Mendelian disorders, with a major gene effect. However, these may not be applicable to disorders with complex inheritance, since importunate candidate variants may filtered-out by the current models Here we propose a methodology that could be applied to disorders with a complex mode of inheritance using linear discriminant analysis (LDA)

Materials and Methods: LDA is useful in statistics for pattern recognition and machine learning in order to find a linear combination of features, which characterizes or separates two or more classes of objects or events [1]. Here we applied LDA to a training set, of 5,718 arbitrarily selected variants from 2,504 individuals from the 1000 Genomes Project [2], comprising variants within the interval chr20:60,343-250,214 (hg19). In order to construct our model, we excluded variants with a correlation ratio higher than 0.9 within the five super population group comprising: AFR (African), AMR (Ad Mixed American) EAS (East Asian), EUR (European) and SAS (South Asian).

Results: As a result, we were able to create a model that unequivocally reclassified all the super-populations from our dataset using 3,277 variants. Given the model, we selected variants which contributed at least with 1% to the classification, resulting in 972 variants that could discriminate the super populations with 92.65% of assertiveness.

Discussion: The methodology presented here proved effective to reclassify individuals from the four super populations even given an arbitrary set of variants from a single chromosome.
Using the complete set of variants from the genome, or those in exons, we could have added more robustness with a more functional spectrum to our predictions with the cost of a large processing time. This methodology has also the potential of application in complex traits with a large number of individuals, helping to reveal the summative effect of a group of variants within the phenotypes.

Conclusion: Given our results, we identified that less than 17% of the original arbitrary variants classify individuals from the 1000 Genomes Project with a high confidence level, showing that this LDA-based method could be applied to variants within exons or the entire genome in assist in the identification of the genetic basis of complex disorders.

References:

[1] Izenman, A. J. (2013). Linear discriminant analysis. In Modern Multivariate Statistical Techniques (pp. 237-280). Springer New York.

[2] 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. Nature, 526(7571), 68-74.