Network Algorithm To Relatedness Analysis (NAToRA)

Thiago Peixoto Leal, Mateus Gouveia, Gilderlanio Santana de Araújo, Maíra R Rodrigues, Marilia Scliar, Eduardo Martin Tarazona Santos

Laboratório de Divergência Genética Humana (LDGH) , Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais

The relatedness can cause population stratification resulting in spurious associations in genome-wide association studies (GWAS) and biases on population structure studies. To correct this problem there are several methods that estimate the relatedness of the samples (Purcell et al 2007 , Thornton et al 2012), giving IBD-sharing probabilities (Identity by descent, ie, the probability that a matching segment of DNA shared by two or more individuals has been inherited from a recent common ancestor). The statistics are composed of pair-wise IBD0, IBD1 and IBD2, which are probability of not share alleles by descent, share one allele by descent and share two alleles by descent respectively. Through IBDs we can calculate the kinship coefficient between two individuals i and j (Φij) by the equation $\Phi ij = \frac{1}{4}\,\mathrm{IBD1}_{ij} + \frac{1}{2}\,\mathrm{IBD2}_{ij}$. The theoretical values of Φij that correspond to the following degree of relatedness between i and j are: 0.5 for self or twins, 0.25 for first degree, 0.125 for second degree, 0.0625 for third degree, 0.03125 for fourth degree and 0 for unrelated. Although there are several methods of estimating kinship , there is no method in the literature that tries to create a population unrelated sample trying to minimize the exclusion of individuals. Using the Graph and Complex Network Theories, NAToRA detects the families in the network and eliminates individuals to create a sample without kinship. The first step is create a Network (a Network (Network N) where the nodes are the individuals and de edges are the kinship coefficient between the nodes. After this step, eliminate all edges between nodes with a value lower than a cut-off value α, ie, what degree of relatedness to be considered (Network N_c). Each cluster in N_c (Connected Component) is a Network Family. As the problem of obtaining a smaller number of individuals to be removed to create a free edge network is a NP-Complete problem, we have implemented a heuristic. The algorithm calculate the node degree centrality and remove the highest degree (ie, the individual with more relatives) until only exist pairs of individuals and edgeless nodes. With the pairs we look for the Network N, calculate the centrality for both nodes and exclude the highest. After the process, the algorithm gives a list of families and the individuals to remove to create a subset without relatives.