

A novel, computationally tractable algorithm flags in big matrices every column associated in any way with others or a dependent variable, with much higher power when columns are linked like mutations in chromosomes

Antezana M.A.

Independent Researcher

marcos.antezana@gmail.com

Key words: synergism, epistasis, non-additivity, additivity, interaction effect, marginal effect, 2-way, 3-way, higher-order effect, linkage disequilibrium, linkage, fine-mapping, coarse-mapping, χ^2 partitioning, mutation, combinations, complex trait, complex disease, broad-sense heritability, model heterogeneity, frequentistic, computational tractability, data mining, big data, large matrices, prime numbers

When a data matrix DM has many independent variables IVs, it is not computationally tractable to assess the association of every distinct IV subset with the dependent variable DV of the DM, because the number of subsets explodes combinatorially as IVs increase. But model selection and correcting for multiple tests is complex even with few IVs.

DMs in genomics will soon summarize millions of mutation markers and genomes. Searching exhaustively in such DMs for markers that alone or synergistically with others are associated with a trait is therefore computationally tractable only for 1- and 2-marker effects. Also population geneticists study mainly 2-marker combinations.

I present a computationally tractable, fully parallelizable Participation in Association Score (PAS) that in a DM with markers detects one by one every column that is strongly associated in any way with others. PAS does not examine column subsets and its computational cost grows linearly with the number of columns, remaining reasonable even when DMs have millions of columns.

PAS exploits how associations of markers in the rows of a DM cause associations of matches in the rows' pairwise comparisons. For every such comparison with a match at a tested column, PAS computes the matches at other columns by modifying the comparison's total matches (scored once per DM), yielding a distribution of conditional matches that reacts diagnostically to the associations of the tested column. Equally computationally tractable is dvPAS that flags DV-associated IVs by perturbing the matches at the DV. P values for the scores are readily obtained by permutation and accurately Sidak-corrected for multiple tests, bypassing model selection. A column's PAS P values PASs for different orders of association under the H_0 and false-positive ones are i.i.d. and readily turned into a single P value. Simulations show that i) PAS and dvPAS generate uniform-(0,1)-distributed type I error in null DMs and ii) detect randomly encountered binary and trinary models of significant n -column association and n -IV association with a binary DV respectively, with power in the order of magnitude of exhaustive evaluation's and false positives that are uniform-(0,1)-distributed or straightforwardly tuned to be so. Power to detect 2-way DV-associated 100-marker+ runs disjunct and not is non-parametrically ultimate but that to detect pure n -column associations and pure n -IV DV associations sinks exponentially as n increases.

a) pure 3-way association			b) pure 4-way association			
0	0	0	0	0	0	0
0	1	1	0	0	1	1
1	0	1	0	1	0	1
1	1	0	1	1	1	0
c) pure 2-way disease model			1	0	0	1
Affecteds	0	0	1	0	1	0
	1	1	1	1	0	0
Controls	0	1	1	1	1	1
	1	0				
d) pure 3-way disease model						
Affecteds	0	0	0			
	0	1	1			
	1	0	1			
	1	1	0			
Controls	0	0	1			
	0	1	0			
	1	0	0			
	1	1	1			

Fig. 1. Pure 3- and 4-way associations among binary markers and related pure 2- and 3-way binary disease models. Panels (a) and (b) show schemes of "pure" 3- and 4-way association of binary markers respectively, that characteristically miss half of the eight and sixteen possible 3- and 4-marker same-row combinations, despite showing all possible lower-order same-row marker combinations. If the 1s and 0s in the leftmost columns in (a,b) are relabelled as "affecteds" and "controls," resp., one obtains the models in (c,d) of "perfect" 2- and pure 3-way disease "causation" in which a set of markers in pure association causes disease and the other set protects against it (i.e., both sets have 100 % "penetrance"), without any differences in the lower-order associations of affecteds and controls, i.e., no marker-frequency differences in (c) and neither marker-frequency nor 2-marker-combination differences in (d). Similarly, a pure 5-way pattern becomes a disease model with pure 4-way causation, etc (not shown)

Important for geneticists, dvPAS power increases about twofold in trinary vs. binary DMs and by orders of magnitude with markers linked like mutations in chromosomes, specially in trinary DMs where furthermore dvPAS fine-maps with highest resolution.