

SUPPORTING MATERIAL

FIGURE S1 BRCA1

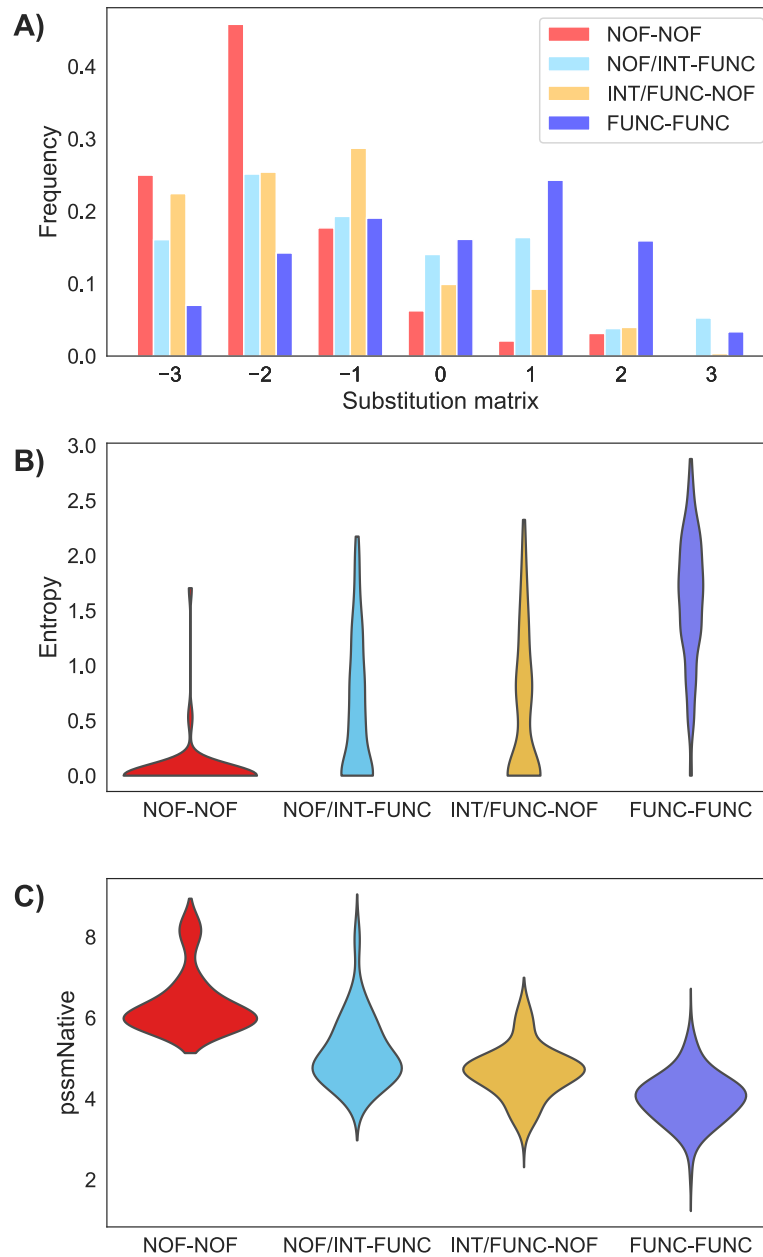


Figure S1. Feature distributions for the outlier populations in Figure 3. The three figures represent value distributions for (A) the substitution matrix (Blosum62) elements, (B) Shannon's entropy, and (C) pssm_{Nat}. In each figure, we display four variant populations from the following regions in Figure 3: FUNC-FUNC (dark blue), NOF-NOF (red), the outliers from the quadrants NOF-FUNC plus INT-FUNC (light blue), and the outliers from the quadrants INT-NOF plus FUNC-NOF (orange). Each region is identified by the functional class of its variants in the HDR and SGE assays; this

functional class is the same for either assay: 'functional' (FUNC), 'intermediate' (INT), and 'non-functional' (NOF).

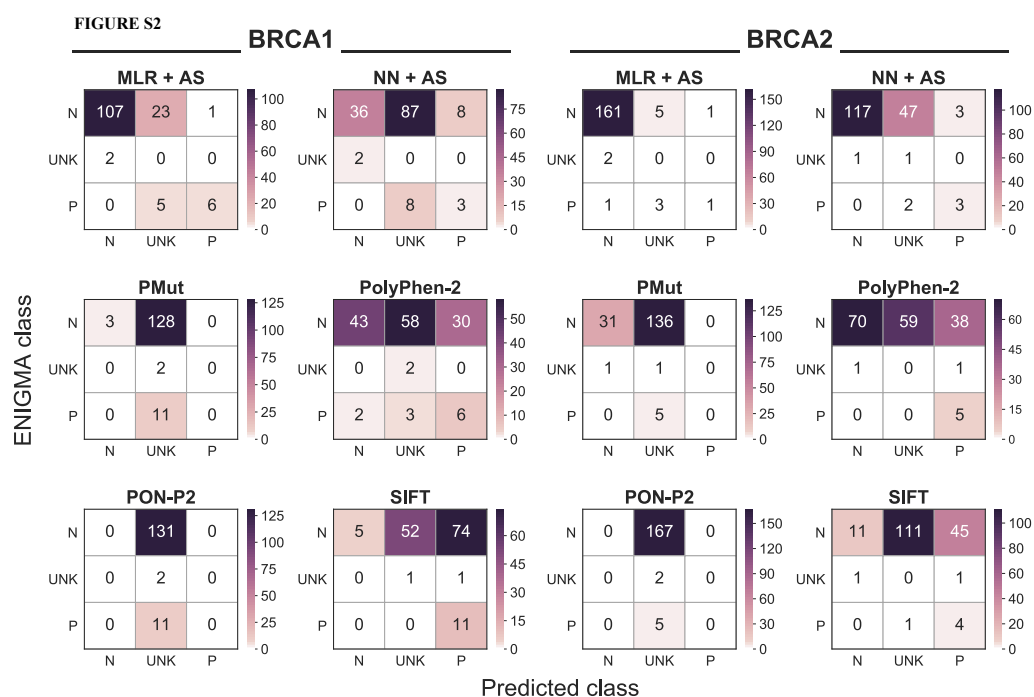


Figure S2. Heatmap of the predictor performances on the CAGI datasets. Each heatmap represents the confusion matrix of a predictor. We provide six heatmaps per protein, two for our predictors (MLR+AS and NN+AS) and four for the general predictors (PolyPhen-2, PON-P2, PMut, and SIFT). In all the plots, the vertical and horizontal axis respectively correspond to the observed (provided by CAGI organizers) and predicted classes, in the 3-class, reduced version of the IARC 5-tier classification. Diagonal and off-diagonal elements correspond to successful and failed predictions, respectively.

NOTE: given the range differences in the predictions, each plot has its color scale.

Table S1. The CAGI predictions. In the separate excel file Table_S1.xlsx, we provide the four list of predictions we submitted to the CAGI experiment. In every list we give, for each of the BRCA1 and BRCA2 variants, the following information: DNA description of the variant; gene name; protein description of the variant; predicted IARC 5-tier class; p; sd; and current IARC 5-tier class according to the ENIGMA consortium.