# SUPPLEMENTARY INFORMATION

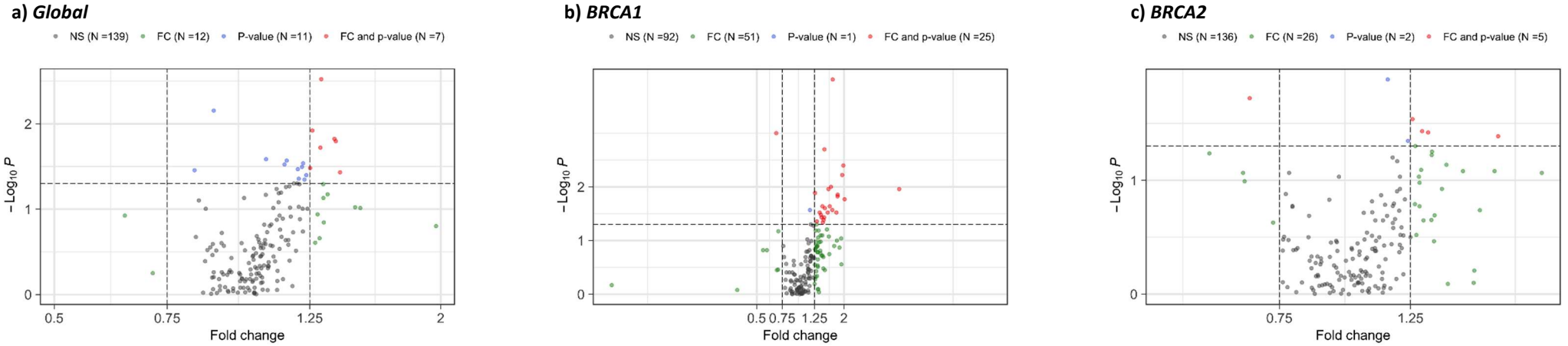
**Metabolomics-driven mutational status prediction in healthy individuals with a family history of hereditary breast and ovarian cancer: the HRRmet study.**

*Bàrbara Roig, Sara Fernández-Castillejo*, Josep Gumà, Joan Badia, Mireia Melé, Mònica Salvat, Montserrat Querol, Raquel Cumeras,* and *Marta Rodríguez-Balada.*

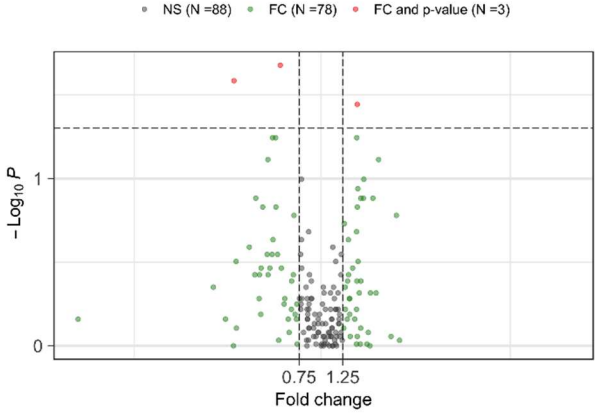
Institut d'Oncologia de la Catalunya Sud (IOCS), Hospital Universitari Sant Joan de Reus (HUSJR), Spain. Institut d'Investigació Sanitària Pere Virgili (IISPV), Reus, Spain. Universitat Rovira i Virgili (URV), Reus, Spain.

## SUPPLEMENTARY FIGURES

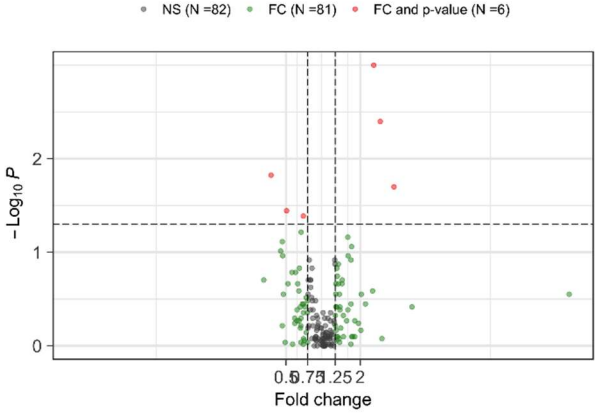
**Supplementary Figure S1**. **Volcano plots showing the statistical significance *vs*. fold change (FC) of all profiled metabolites (n=169)**. **a)** Global plot; **b)** *BRCA1* plot; **c)** *BRCA2* plot; **d)** *PALB2* plot; **e)** *ATM* plot; **f)** *CHEK2* plot; **g)** RAD51 plot; **h)** High-penetrance plot; **i)** Moderate-penetrance plot. Grey dots indicate those metabolites not meeting the FC (FC>1.25 or FC<0.75) nor p-value (p<0.05) criteria. Green dots indicate metabolites meeting only the FC criterium, whilsts blue dots those meeting only the p-value criterium. Red dots indicate metabolites meeting both the FC and the p-value criteria.



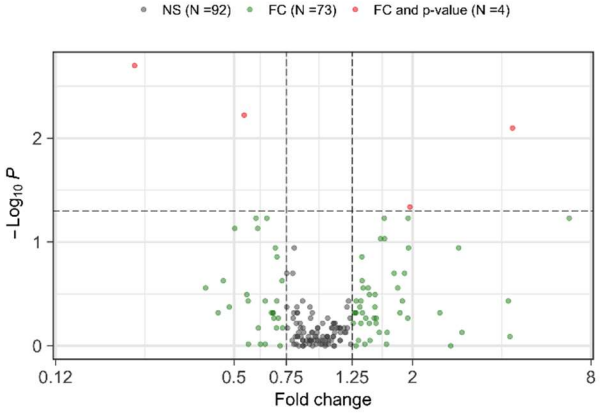
**d) *PALB2***



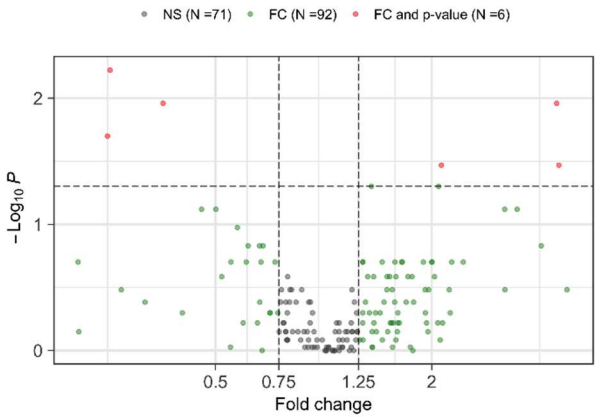

**e) *ATM***



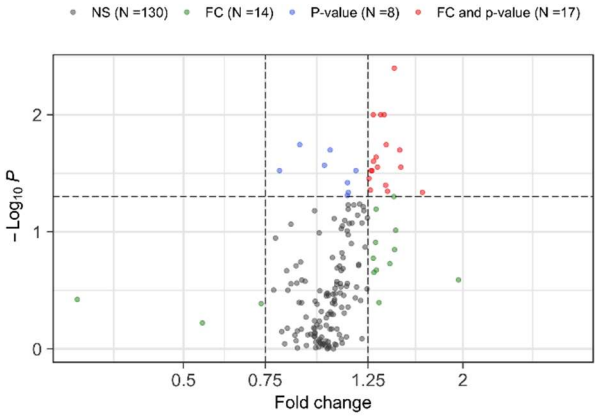

**f) *CHEK2***



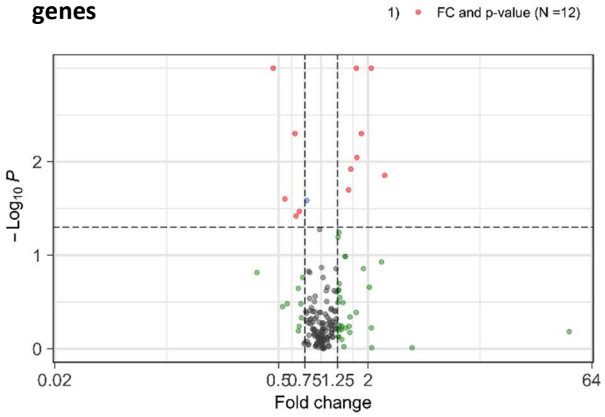

**g) *RAD51***



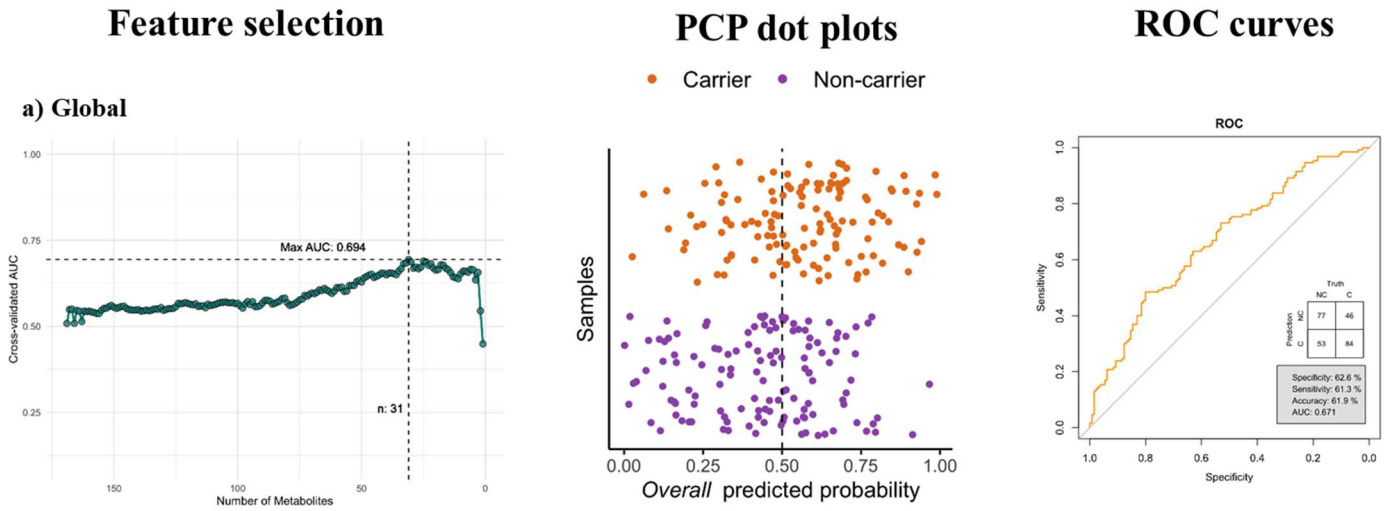

**h) High-penetrance genes**




**i) Moderate-penetrance genes**



**Supplementary Figure S2. Machine-learning approach based on a linear Support Vector Machine (SVM). a)** Global model; **b)** *PALB2* model; **c)** *ATM* model; **d)** *CHEK2* model; **e)** High-penetrance model; **f)** Moderate-penetrance model. Linear SVM was used to compute the predicted class probability for carriers *vs* non-carriers for each subset of samples. First, in the feature selection step, the recursive feature elimination (RFE) method was chosen with 3-fold cross-validation for the iterative removal of the least important features (metabolites) to enhance model simplicity. Maximum predictive performance measured by the maximal Area Under the Receiver-Operating Characteristic curve (maxAUC) with the minimal number of metabolites, was used to select the final metabolites set. VI score for each selected metabolite was calculated using the coefficients (weights) calculated by the SVM model. Individual AUC for each selected metabolite was also calculated using its own abundances. Second, the final predictive model was built with the VI metabolites selected in the feature selection step, using a 3-fold cross-validation SVM. The Receiver-Operating Characteristic (ROC) curve and the Predicted Carrier Probability (PCP) dot plot were constructed using the predicted SVM scores. The confusion matrix, specificity, sensitivity and accuracy were calculated using a 0.5 probability threshold.



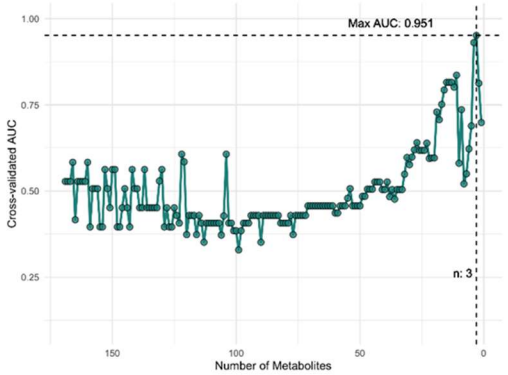
## Feature selection

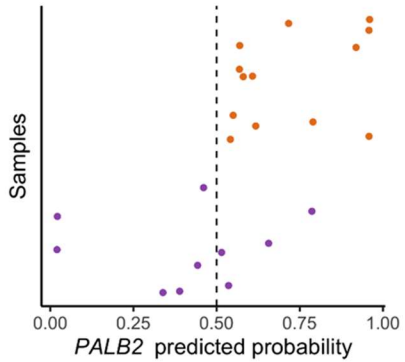
## PCP dot plots

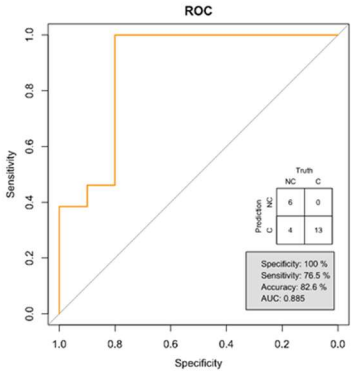
## ROC curves

Carrier Non-carrier

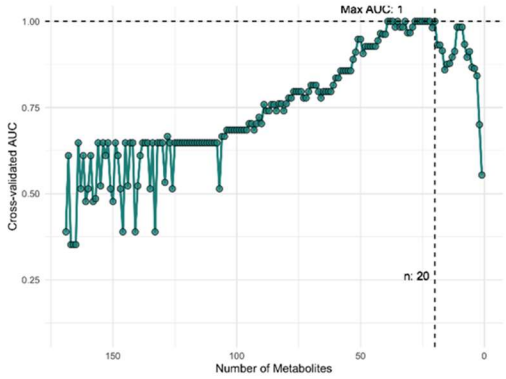
### b) PALB2

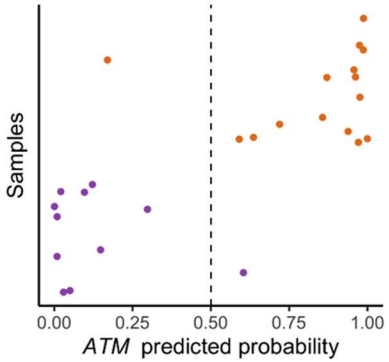


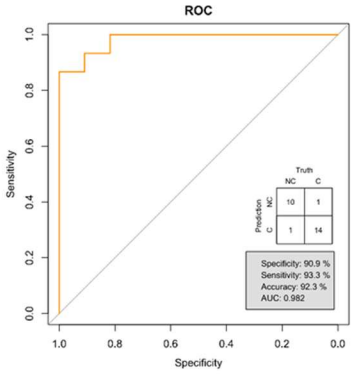



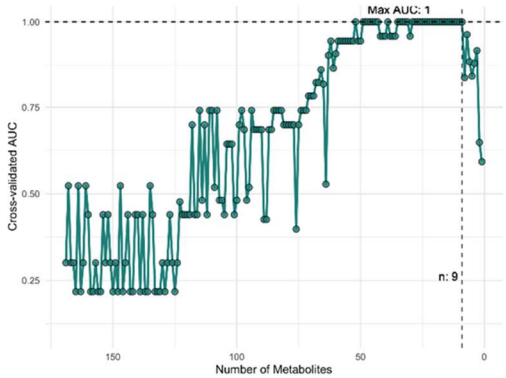

### c) ATM

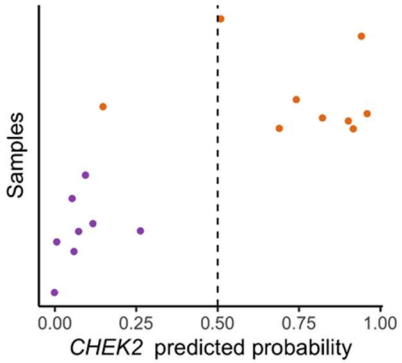


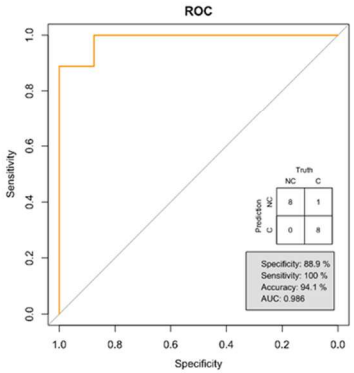




### d) CHEK2



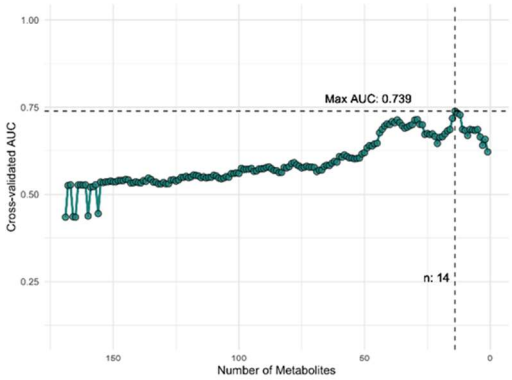
## Feature selection

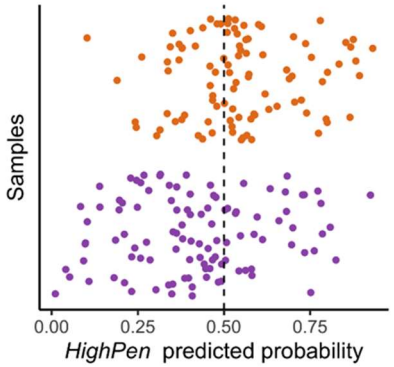
## PCP dot plots

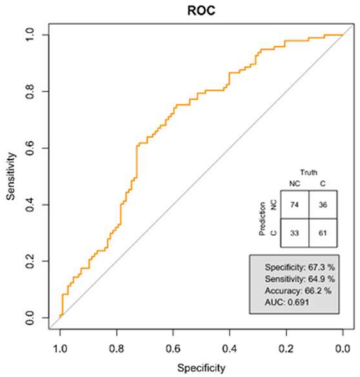
## ROC curves

Carrier Non-carrier

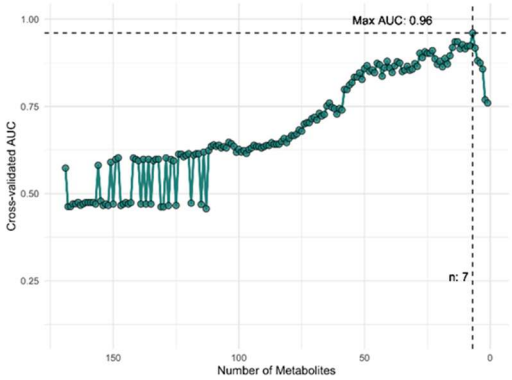
### e) High penetrance

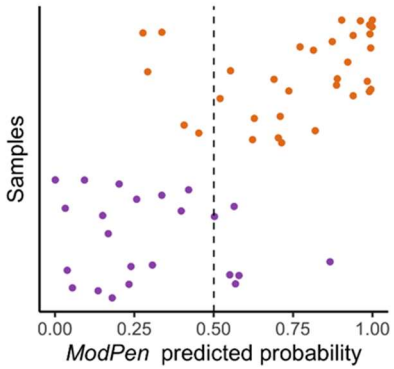


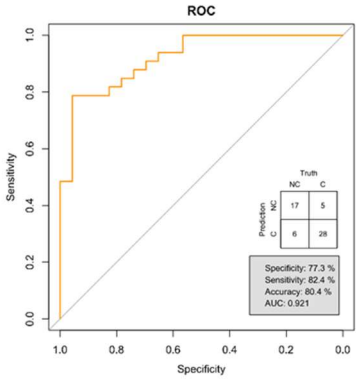




### f) Moderate penetrance



**SUPPLEMENTARY TABLES**

**Supplementary Table S1. Peak intensities (abundances) of metabolites significantly present between carrier and non-carrier subpopulations.** Only those metabolites that were significant ($p<0.05$) in any of the comparisons performed are included in this table. Likewise, only data that were significant is included in the table. Peak intensities are expressd as mean and standard deviations. Abbreviations: SD, Standard Deviation.

**Supplementary Table S2. Fold changes (FC) of all the metabolites profiled in this study.** FC were calculated as the ratio of each metabolite abundance in carriers/non carriers. $FC>1.25$ and $FC<0.75$ cutoff limits were selected arbitrarily, considering that since participants were healthy individuals, no greater FCs were expected. Those values that met the FC criteria are indicated in blue. Those values that were significant (p-value $<0.05$) are indicated in bold. Those metabolites that ere selected as VI metabolite in the predictive models are highlighted in green.