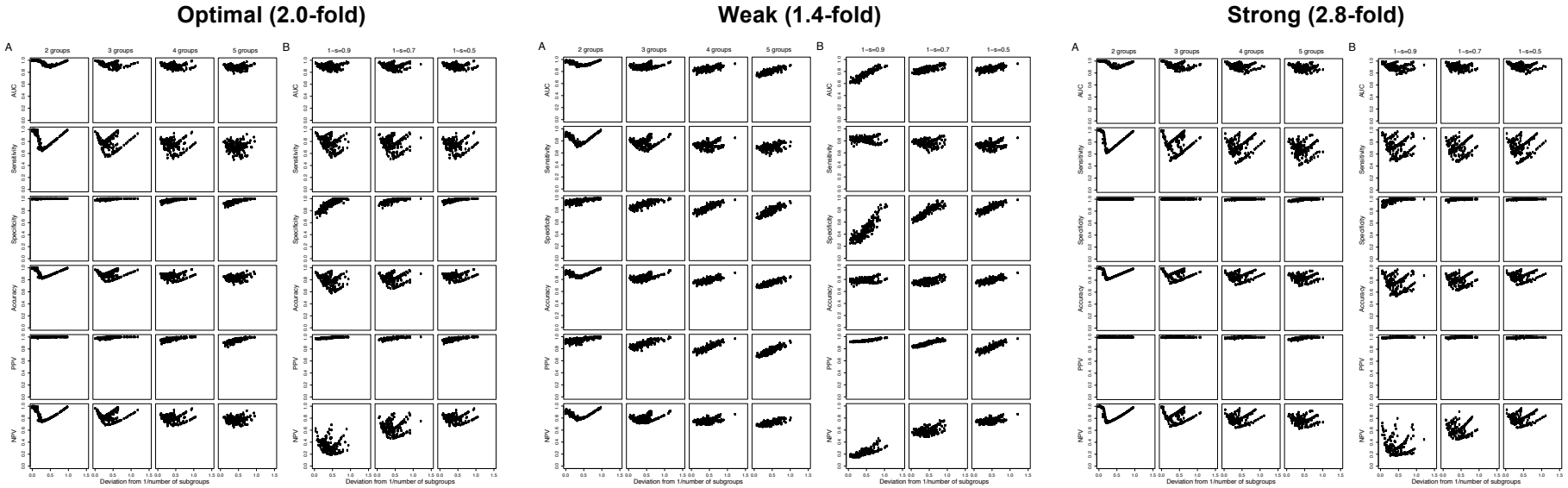
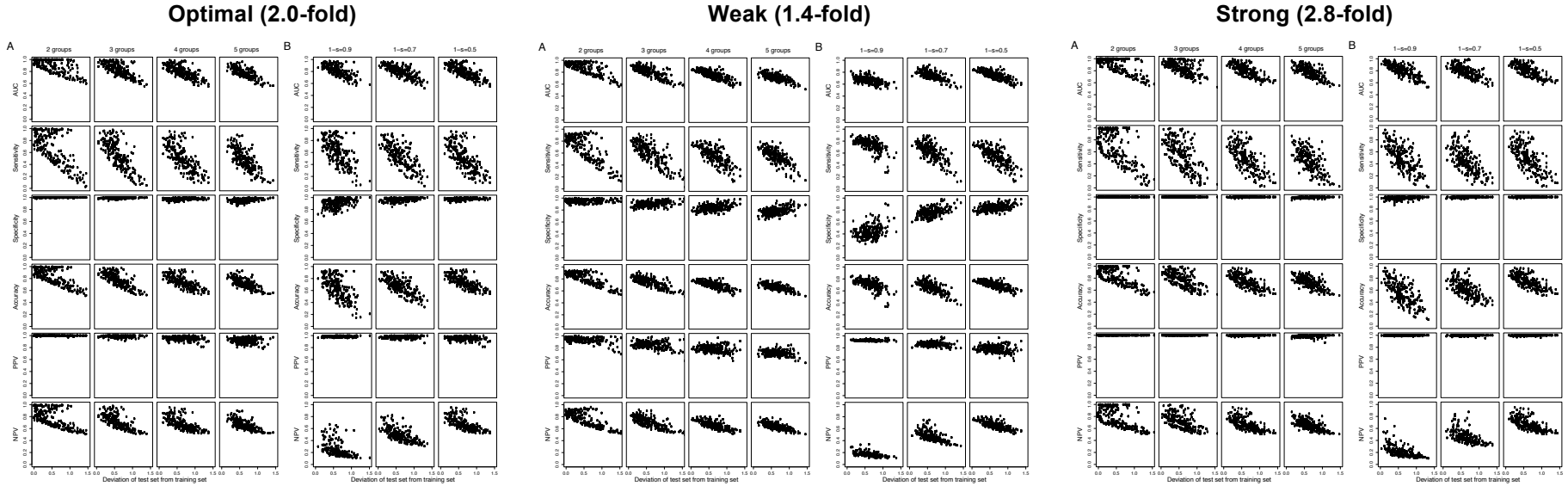
Supplementary Figure S1: Impact of unevenly distributed resistance mechanisms between training and test sets on signature performance.



Comparative impact of multiple unevenly distributed optimal (2.0-fold), weak (1.4-fold) and strong (2.8-fold) resistance mechanisms with identical prevalence in training and test sets on the performance of predictive gene signatures. Perturbed datasets in which *s% (s%=5%, 10%, 20%, 30%, 40% or 50%)* of the cases were designated to be therapy sensitive were generated. Within the resistant *1-s%* cases, the cases were allocated randomly into *n (n=2, 3, 4, 5)* groups of resistance mechanisms. For each n^{th} resistance mechanism, 100 genes were randomly selected as the "true" gene expression changes and were spiked-in by *v (v=1*, left; *v=0.5*, middle; *v=1.5*, right*)*. Classification was performed using diagonal linear discriminant analysis (DLDA). For each combination of *s* and *n,* we repeated the spiking and classification 200 times. The performance of the predictive gene signature for each of the 200 repeats where each data point represents the median of 50 Monte-Carlo Cross Validation (MCCV) repeats is shown. The performance of the predictive gene signature was measured by the area under curve (AUC) of receiver operating characteristic (ROC) curves, sensitivity, specificity, accuracy, positive predictive value (PPV), negative predictive value (NPV), and the proportion of spiked-in genes recovered in the inferred signatures. (A) Within each row, the performance of the predictive gene signature is plotted against deviation of the sizes of the subgroups from *1/n*, calculated as $\sum_{i=2}^{n} |f_i - \frac{1}{n}|$, where f_i is the size of the i^{th} subgroup, for (from left) *n=2* (labeled "2 groups"), *n=3* (labeled "3 groups"), *n=4* (labeled "4 groups") and *n=5* (labeled "5 groups"). (B) Within each row, the performance of the predictive gene signature is plotted against deviation of the sizes of the subgroups from *1/n*, calculated as $\sum_{i=2}^{n} |f_i - \frac{1}{n}|$, where f_i is the size of the i^{th} subgroup, for (from left) 1-*s%=0.9*, 1-*s%=0.7 and* 1-*s%=0.5.*

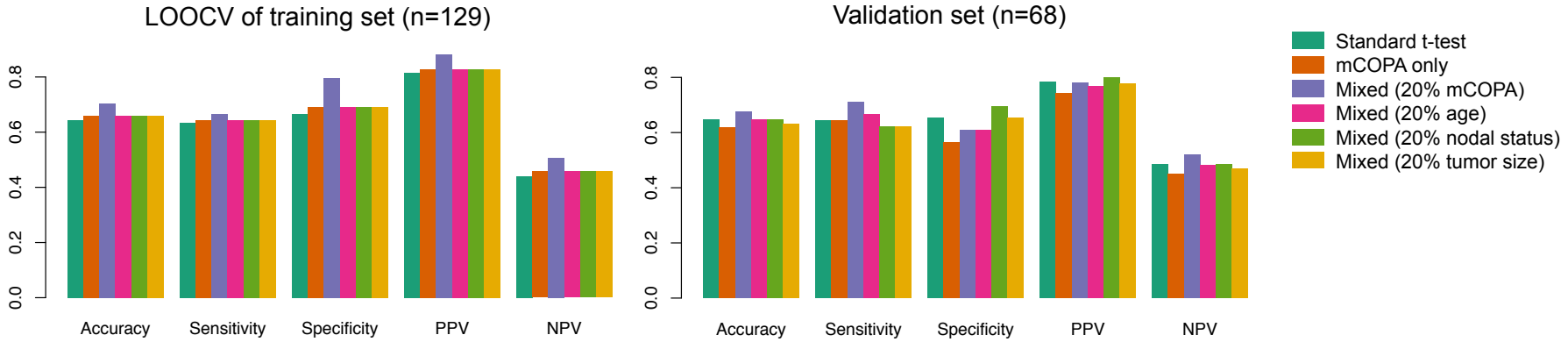
Supplementary Figure S2: Impact of different distributions of resistance mechanisms in training and test sets on signature performance.



Comparative impact of multiple unevenly distributed optimal (2.0-fold), weak (1.4-fold) and strong (2.8-fold) resistance mechanisms with random and independent prevalence in training and test sets on the performance of the predictive gene signatures. In both, test and training set, the total proportion of resistant cases is identical. Perturbed datasets in which *s% (s%=5%, 10%, 20%, 30%, 40% or 50%)* of the cases were designated to be therapy sensitive were generated. Within the resistant *1-s%* cases, the cases were allocated randomly into *n (n=2, 3, 4, 5)* groups of resistance mechanisms and the case allocation for training and test datasets was performed independently. Furthermore, for each n^{th} resistance mechanism, 100 genes were randomly selected as the "true" gene expression changes and were spiked-in by *v (v=1*, left; *v=0.5*, middle; *v=1.5*, right*)*. Classification was performed using diagonal linear discriminant analysis (DLDA). For each combination of *s* and *n,* we repeated the spiking and classification 200 times. The performance of the predictive gene signature for each of the 200 repeats where each data point represents the median of 50 Monte-Carlo Cross Validation (MCCV) repeats is shown. The performance of the predictive gene signature was measured by the area under curve (AUC) of receiver operating characteristic (ROC) curves, sensitivity, specificity, accuracy, positive predictive value (PPV), negative predictive value (NPV), and the proportion of spiked-in genes recovered in the inferred signatures. (A) Within each row, the performance of the predictive gene signature is plotted against deviation of the sizes of the distinct resistance mechanism groups in the test dataset from those in the training dataset, calculated as $\sum_{i=2}^{n} |f_{i,test} - f_{i,train}|$, where $f_{i,test}$ is the size of the i^{th} subgroup in the test set and $f_{i,train}$ is the size of the i^{th} subgroup in the training set, for (from left) *n=2* (labeled "2 groups"), *n=3* (labeled "3 groups"), *n=4* (labeled "4 groups") and *n=5* (labeled "5 groups"). (B) Within each row, the performance of the predictive gene signature is plotted against deviation of the sizes of the distinct resistance mechanism groups in the test dataset from those in the training dataset, calculated as $\sum_{i=2}^{n} |f_{i,test} - f_{i,train}|$, where $f_{i,test}$ is the size of the i^{th} subgroup in the test set and $f_{i,train}$ is the size of the i^{th} subgroup in the training set, for (from left) 1-*s%=0.9*, 1-*s%=0.7 and* 1-*s%=0.5.*

Supplementary Figure S3: **Impact of cohort sub-stratification on signature performance.**

	LOOCV of training set (n=129)						Validation set (n=68)					
Performance measure	**Standard t-test**	**mCOPA only**	**Mixed (20% mCOPA)**	**Mixed (20% age)**	**Mixed (20% nodal status)**	**Mixed (20% tumor size)**	**Standard t-test**	**mCOPA only**	**Mixed (20% mCOPA)**	**Mixed (20% age)**	**Mixed (20% nodal status)**	**Mixed (20% tumor size)**
Accuracy	0.643	0.659	0.705	0.659	0.659	0.659	0.647	0.618	0.676	0.647	0.647	0.632
Sensitivity	0.633	0.644	0.667	0.644	0.644	0.644	0.644	0.644	0.711	0.667	0.622	0.622
Specificity	0.667	0.692	0.795	0.692	0.692	0.692	0.652	0.565	0.609	0.609	0.696	0.652
PPV	0.814	0.829	0.882	0.829	0.829	0.829	0.784	0.744	0.780	0.769	0.800	0.778
NPV	0.441	0.458	0.508	0.458	0.458	0.458	0.484	0.448	0.519	0.483	0.485	0.469



Impact of sub-stratification of chemotherapy-resistant ER-negative breast cancers, based on the expression of outliers or on gene expression patterns associated with established clinical parameters, on predictive signature performance in actual (non-bioinformatically manipulated) breast cancer datasets. Predictive signatures were generated using a standard linear t-test ('standard t-tests'), a modified Cancer Outlier Profiling Analysis (mCOPA), a mixed mCOPA (20%) and t-test approach (80%; 'Mixed (20% mCOPA)'), or a mixed approach for clinical parameters, including a mixed standard t-test (80%) and age at diagnosis-related signatures (20%; 'Mixed (20% age)'), a mixed standard t-test (80%) and nodal status-related signatures (20%; 'Mixed (20% nodal status)'), or a mixed standard t-test (80%) and tumor size-related signatures (20%; 'Mixed (20% tumor size)') for feature selection in a training set of 129 taxane-anthracycline-based chemotherapy-resistant ER-negative breast cancers. Validation of the predictive signatures generated was performed by leave-one-out cross-validation (LOOCV) of the training set (n=129), and the accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) are shown. Predictive signatures generated in the training set were validated in an independent dataset of taxane-anthracycline-resistant ER-negative breast cancers (n=68), and the accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of these signatures are shown.