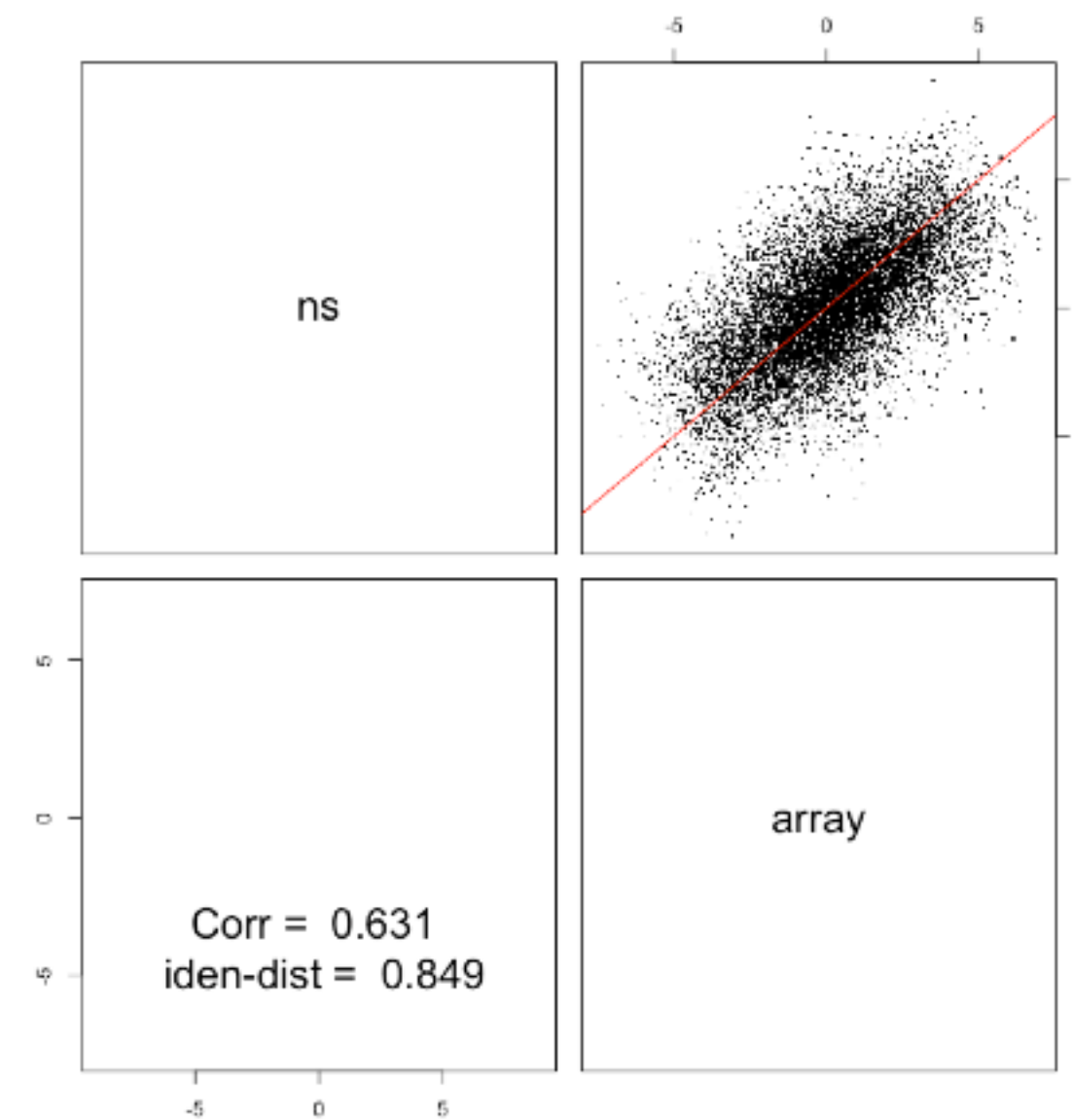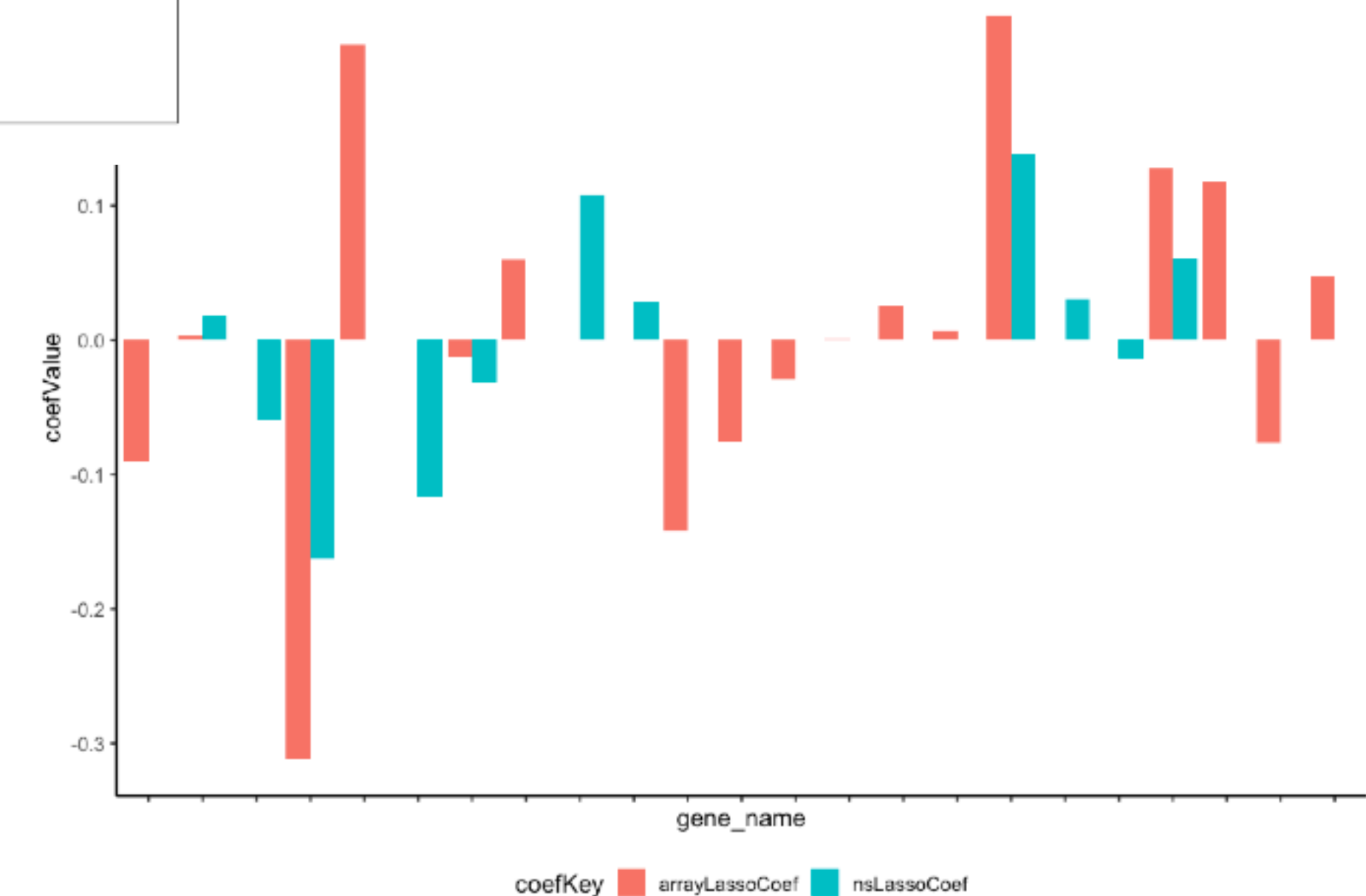# CPOP framework: weighted variable selection

1. Perform a **weighted Lasso** by placing higher weights on features closer to the identity line

2. Perform a **Ridge regression** and only retain those features with coefficients similar to each other
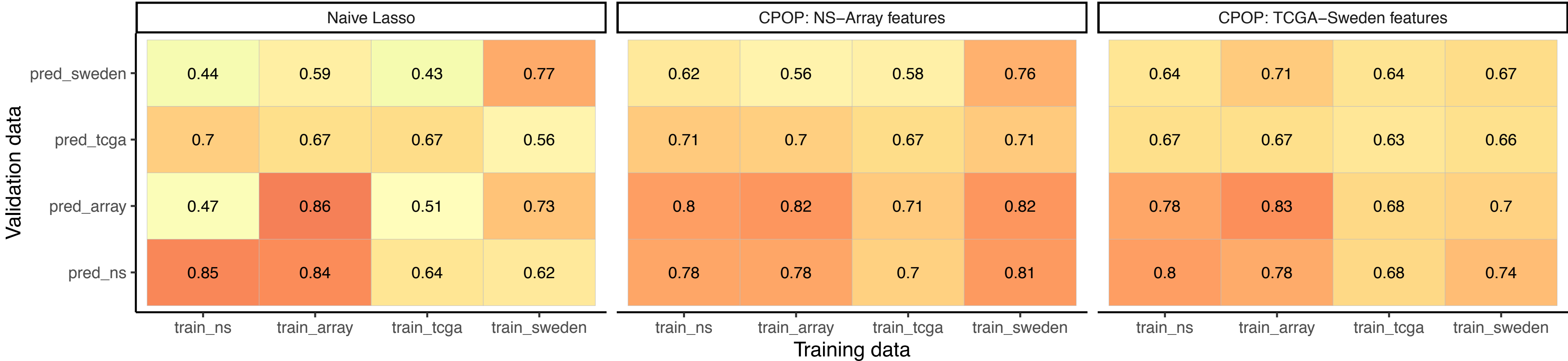
$$Z_1 \approx Z_2$$

$$\hat{\beta}_1 \approx \hat{\beta}_2$$

CPOP results

1. Predictive performance of CPOP matches that of re-substitution.

2. Smaller identity distance between predicted values

F1 classification statistic under various models

Validation data	Naive Lasso				CPOP: NS−Array features				CPOP: TCGA−Sweden features			
	train_ns	train_array	train_tcga	train_sweden	train_ns	train_array	train_tcga	train_sweden	train_ns	train_array	train_tcga	train_sweden
pred_sweden	0.44	0.59	0.43	0.77	0.62	0.56	0.58	0.76	0.64	0.71	0.64	0.67
pred_tcga	0.7	0.67	0.67	0.56	0.71	0.7	0.67	0.71	0.67	0.67	0.63	0.66
pred_array	0.47	0.86	0.51	0.73	0.8	0.82	0.71	0.82	0.78	0.83	0.68	0.7
pred_ns	0.85	0.84	0.64	0.62	0.78	0.78	0.7	0.81	0.8	0.78	0.68	0.74

Training data

Identity distance between predicted values under various models

Validation data	Naive Lasso				CPOP: NS−Array features				CPOP: TCGA−Sweden features			
	train_ns	train_array	train_tcga	train_sweden	train_ns	train_array	train_tcga	train_sweden	train_ns	train_array	train_tcga	train_sweden
pred_sweden	0.08	0.07	0.09	0	0.04	0.04	0.03	0	0.04	0.04	0.02	0
pred_tcga	0.06	0.11	0	0.09	0.06	0.07	0	0.03	0.05	0.06	0	0.01
pred_array	0.16	0	0.19	0.12	0.01	0	0.07	0.04	0.04	0	0.09	0.07
pred_ns	0	0.06	0.11	0.22	0	0.01	0.08	0.06	0	0.06	0.08	0.07

Training data