Supporting Information for:

Automated Protocol for Large-Scale Modeling of Gene Expression Data

Michelle Lynn Hall,[*,†,¶] David Calkins,[‡] and Woody B. Sherman[†]

†Schrödinger, Inc., 222 Third Street, Cambridge, MA 02143

‡Schrödinger, Inc., 101 SW Main St #1300, Portland, OR 97204

¶Current address: Moderna Therapeutics, 200 Technology Square, Cambridge, MA 02139

[*] ***E-mail: michelle.lynn.hall@gmail.com***

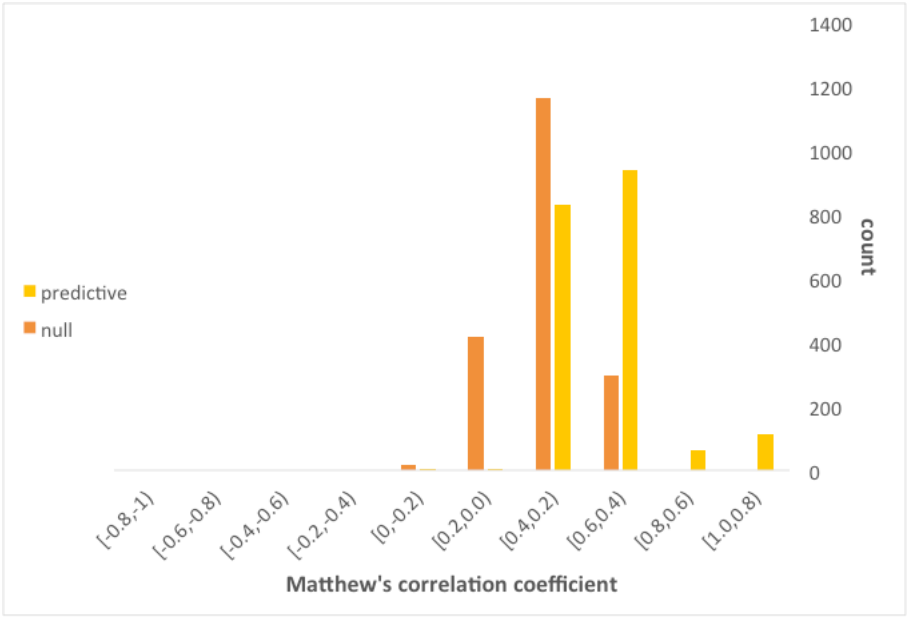

Figure S1: Distribution of Matthew's Correlation Coefficients across training sets for the predictive and null models

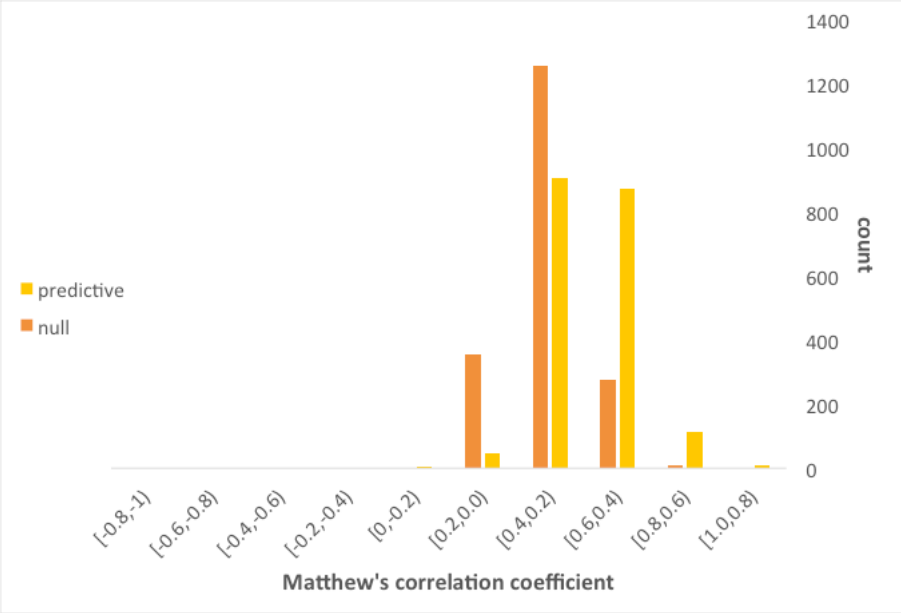

Figure S2: Distribution of Matthew's Correlation Coefficients across test sets for the predictive and null models

Table S1: Matthew's Correlation Coefficient statistics for training and test sets of both predictive and null models

	predictive models		null models	
	training set	test set	training set	test set
minimum	-0.04	-0.03	-0.07	0.03
maximum	1.00	1.00	0.56	0.80
average	0.45	0.41	0.29	0.29
median	0.42	0.40	0.30	0.29
standard deviation	0.15	0.12	0.11	0.10

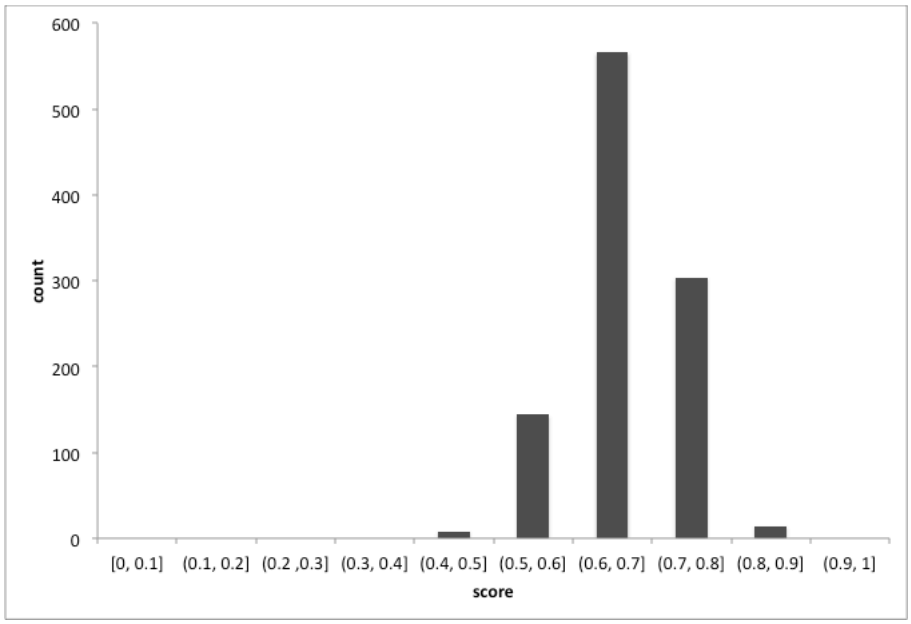

Figure S3: Distribution of *score* for the predictive models when generated based upon two categories only (i.e., affected or not) instead of creating two, two-category models [i.e., (a) up-regulated or not and (b) down-regulated or not]. Compared to the plots shown in Figure 4, we see a slight degradation in score. In particular, the average score here is only 0.67, compared to 0.71 for the scores shown in Figure 4.