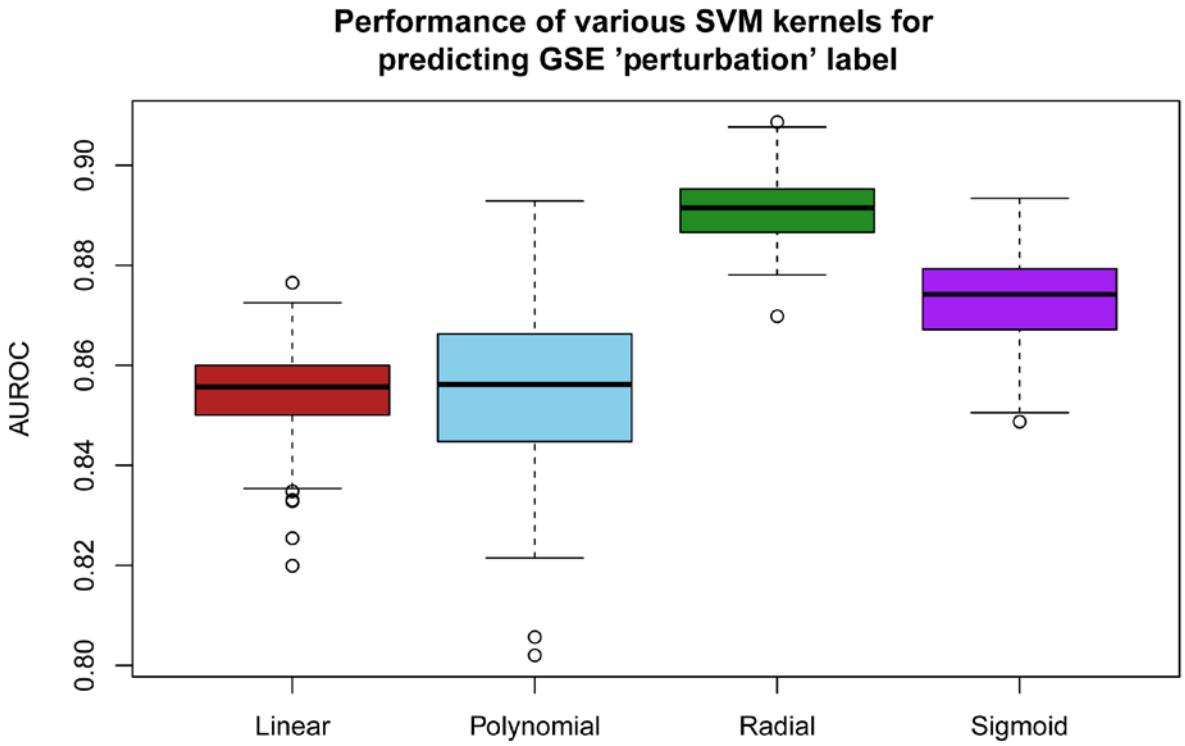
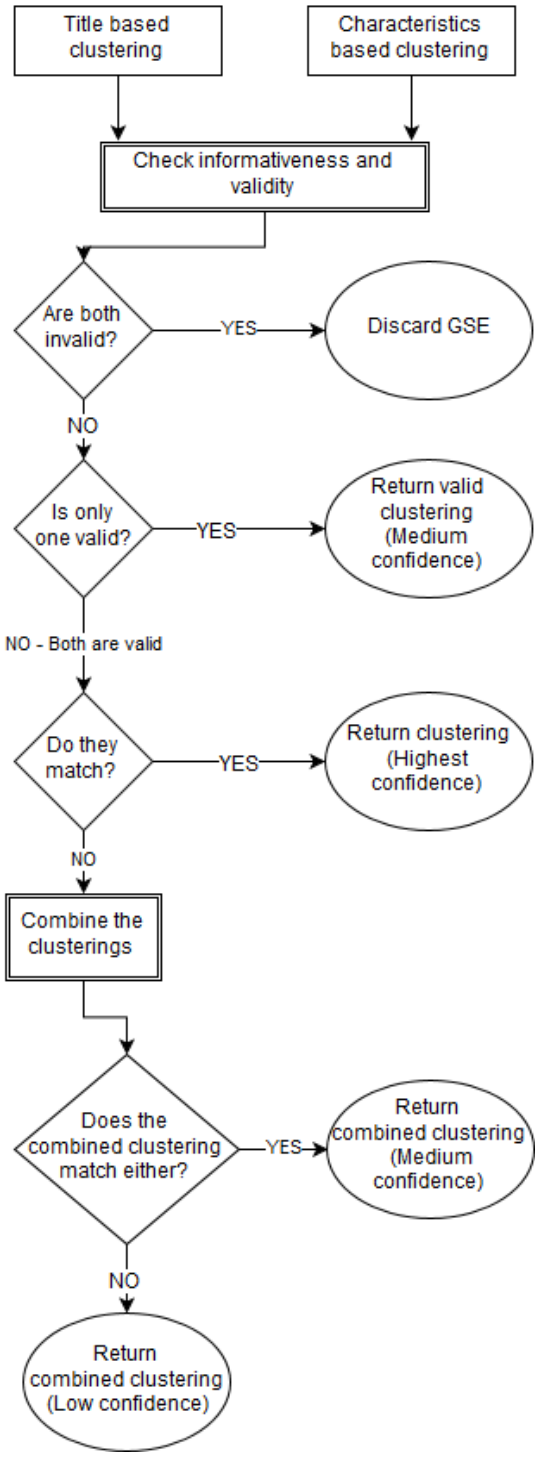
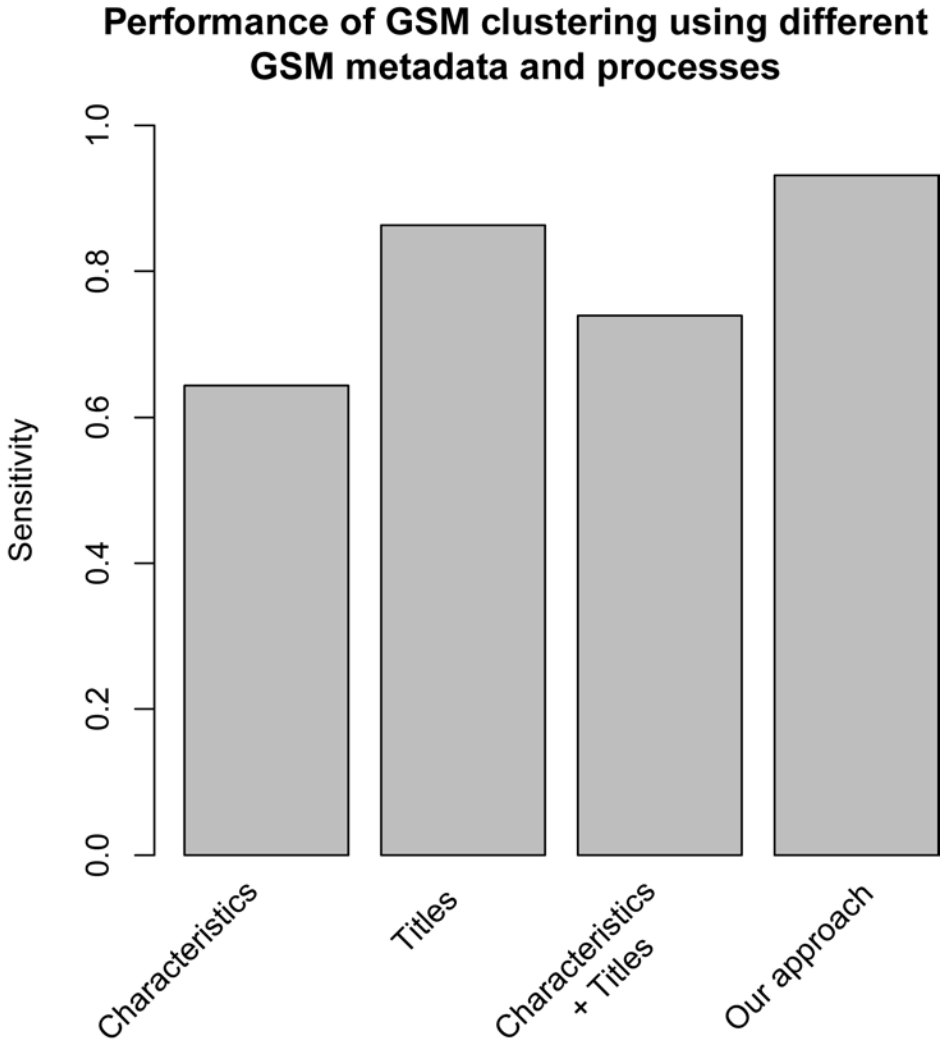
Supplementary Figure 1: Comparison of the performance of different SVM kernels for predicting GSE 'Perturbation' label based on the manually curated training set of 277 GSE IDs. Shown are boxplots of the Area Under the Receiver Operating Characteristic (AUROC) curve from 100 repetitions of 10-fold cross-validation.



Supplementary Figure 2: The logic flow for assessing the most valid clustering of GSM samples. This schematic diagram shows the decision making process during the multi-stage clustering procedure that combines information from the GSM titles and characteristics. Informativeness and validity means that there is more than 1 cluster in the GSE and that there are fewer than N clusters, where N is the number of GSM in the GSE.



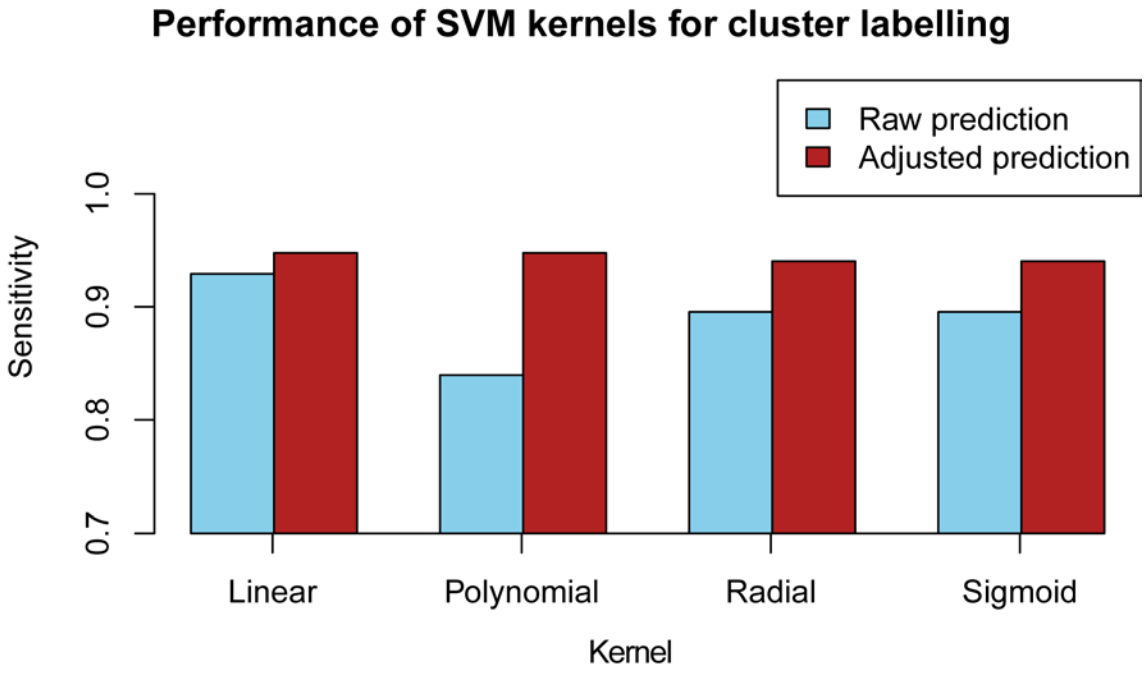
Supplementary Figure 3: Comparing the performance of clustering using GSM titles and characteristics. Shown is the relative sensitivity of different clustering methods, using GSM characteristics only, GSM titles only, a simple concatenation of GSM characteristics and titles and our multi-stage clustering approach.



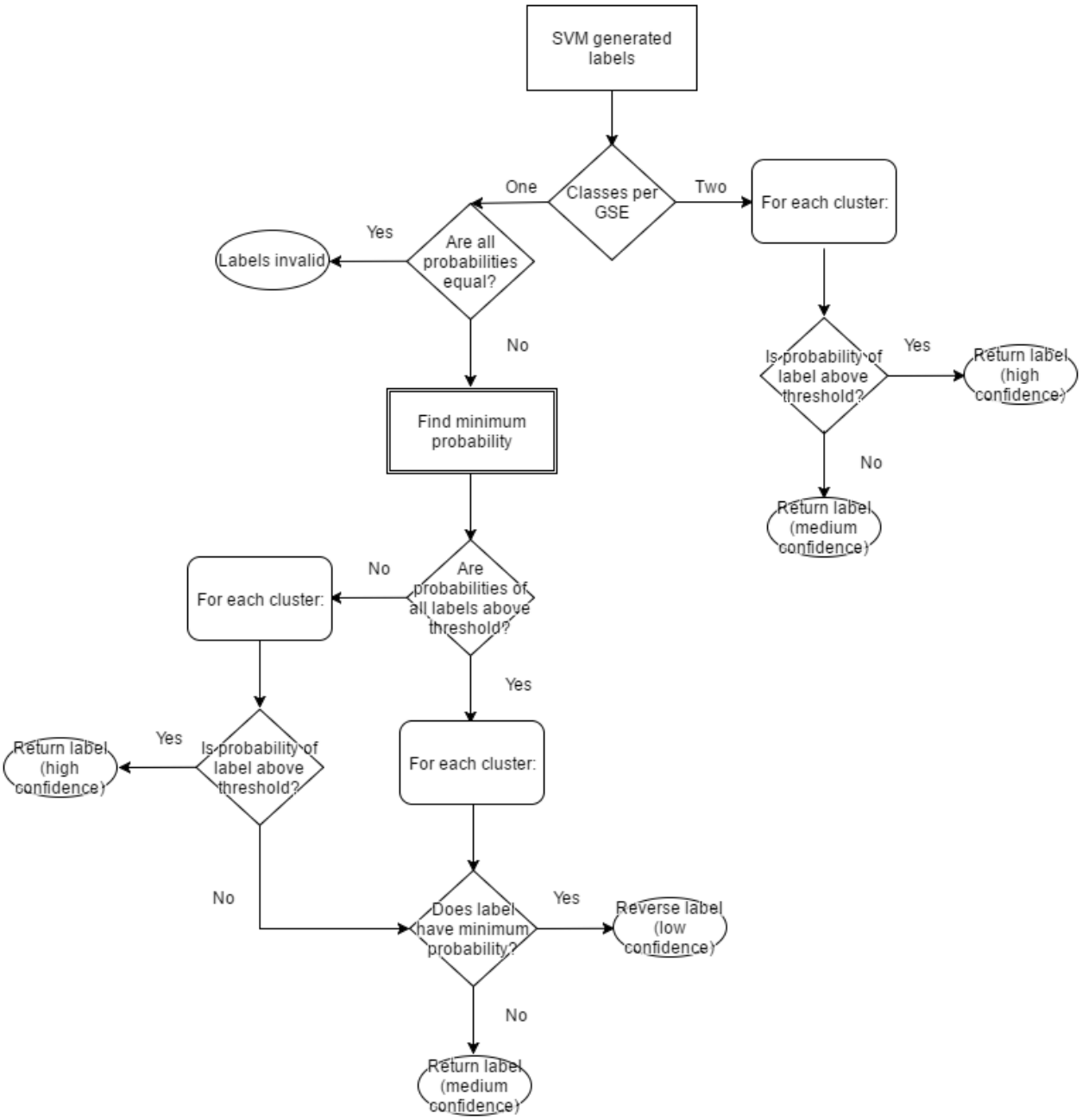
Supplementary Figure 4: Sample titles from GSE41674. Title based clustering was not able to correctly cluster this GSE, whereas GEOracle's multistage clustering approach could, by utilizing the information in the GSM characteristics.

GSM	TITLE	gender	strain	tissue	stage	genotype.variation
GSM1022194	ZJGXZ3_E15-4	male	b6/129sv	embryonic heart	15.5 dpc	zic3 +/y
GSM1022195	ZJGXZ3_E15-6	male	b6/129sv	embryonic heart	15.5 dpc	zic3 +/y
GSM1022196	WT-1	male	b6/129sv	embryonic heart	15.5 dpc	zic3 +/y
GSM1022197	WT-2	male	b6/129sv	embryonic heart	15.5 dpc	zic3 +/y
GSM1022198	WT-5	male	b6/129sv	embryonic heart	15.5 dpc	zic3 +/y
GSM1022199	ZJGXZ3_678-4	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y
GSM1022200	ZJGXZ3_678-5	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y
GSM1022201	ZJGXZ3_628-1	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y
GSM1022202	680-5	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y
GSM1022203	741-3	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y
GSM1022204	741-4	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y
GSM1022205	ZJGXZ3_628-4	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y; sox2-cre
GSM1022206	ZJGXZ3_754-2	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y; sox2-cre
GSM1022207	ZJGXZ3_754-4	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y; sox2-cre
GSM1022208	913-1	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y; sox2-cre
GSM1022209	882-1	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y; sox2-cre
GSM1022210	882-2	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y; sox2-cre
GSM1022211	ZJGXZ3_1191-2	male	b6/129sv	embryonic heart	15.5 dpc	zic3 -/y
GSM1022212	ZJGXZ3_1191-3	male	b6/129sv	embryonic heart	15.5 dpc	zic3 -/y
GSM1022213	ZJGXZ3_1235-8	male	b6/129sv	embryonic heart	15.5 dpc	zic3 -/y
GSM1022214	1322-3	male	b6/129sv	embryonic heart	15.5 dpc	zic3 -/y
GSM1022215	1493-1	male	b6/129sv	embryonic heart	15.5 dpc	zic3 -/y

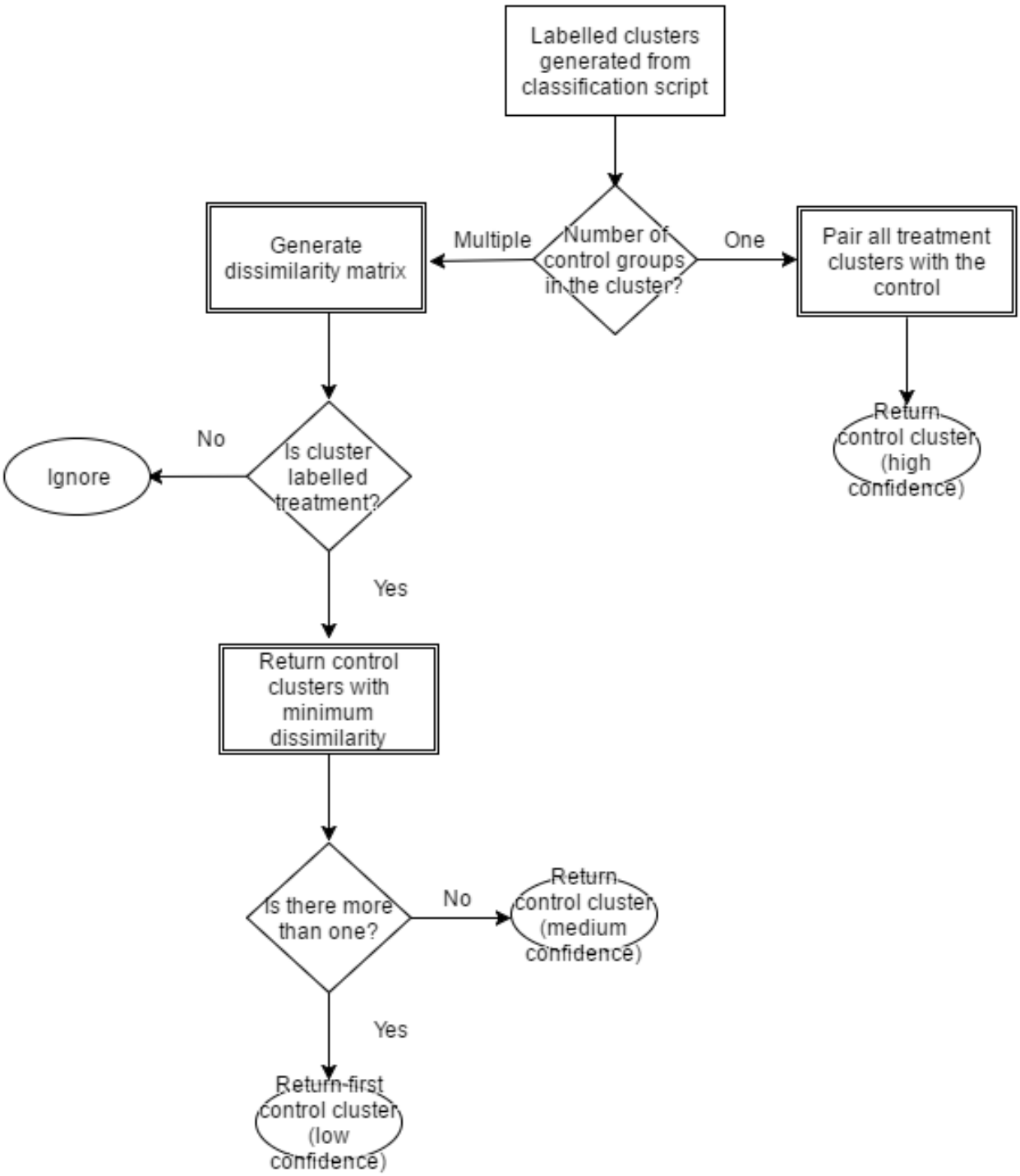
Supplementary Figure 5: Comparing the performance of different SVM kernels to predict the label of GSM clusters ('perturbation' vs 'control'). Sensitivity is calculated as the fraction of GSE for which the GEOracle output perfectly matches the manually annotated set of 73 GSE. Shown is sensitivity calculated on the raw label predictions (blue) and after cluster label adjustment (red).



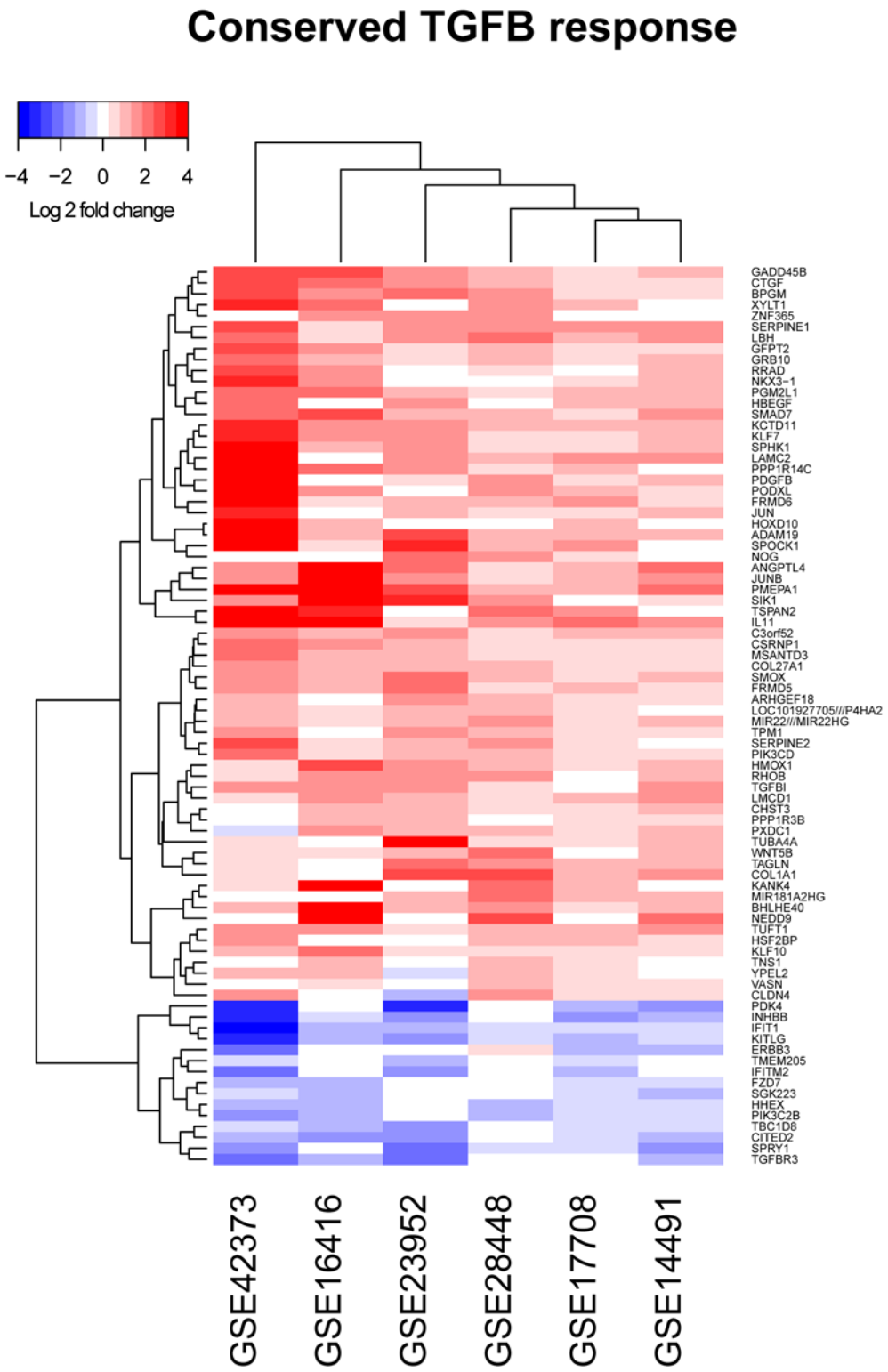
Supplementary Figure 6: The logic flow for assessing the most valid label for a cluster of GSM. This schematic diagram shows the decision making process for fixing labels ('perturbation' or 'control') predicted by the SVM based on textual features. This process is particularly important when only one cluster label is generated for every cluster in a GSE.



Supplementary Figure 7: The logic flow for pairing labelled clusters. This schematic diagram shows the decision making process for matching a 'perturbation' cluster which its closest 'control' cluster. This can be non-trivial when multiple 'control' clusters exist within a GSE. Dissimilarity is Gower distance.



Supplementary Figure 8: A heat map showing the discovered conserved response to TGFB stimulation in human cells. This plot is generated in case study 3.1**.**



Supplementary Figure 9: Mouse heart causal gene regulatory network. A) Overview of the network, showing 23,347 edges between 9152 genes. **B)** Zoom view, showing the directed and signed (activating vs inhibiting) information contained in a subset of the network.