

Supplementary Materials: Impact of Class Imbalance Correction on Biological Feature Selection in Pancreatic Ductal Adenocarcinoma Methylation Data

Paige Keller¹, Zachary Caterer^{2,3,4}, Kaia Hansen¹, Nijhum Paul⁵, Amanda Warren⁶, Samuel T. Morehead⁷, Rick J. Jansen⁸ and Rahul Gomes^{1,*}

Supplementary Information

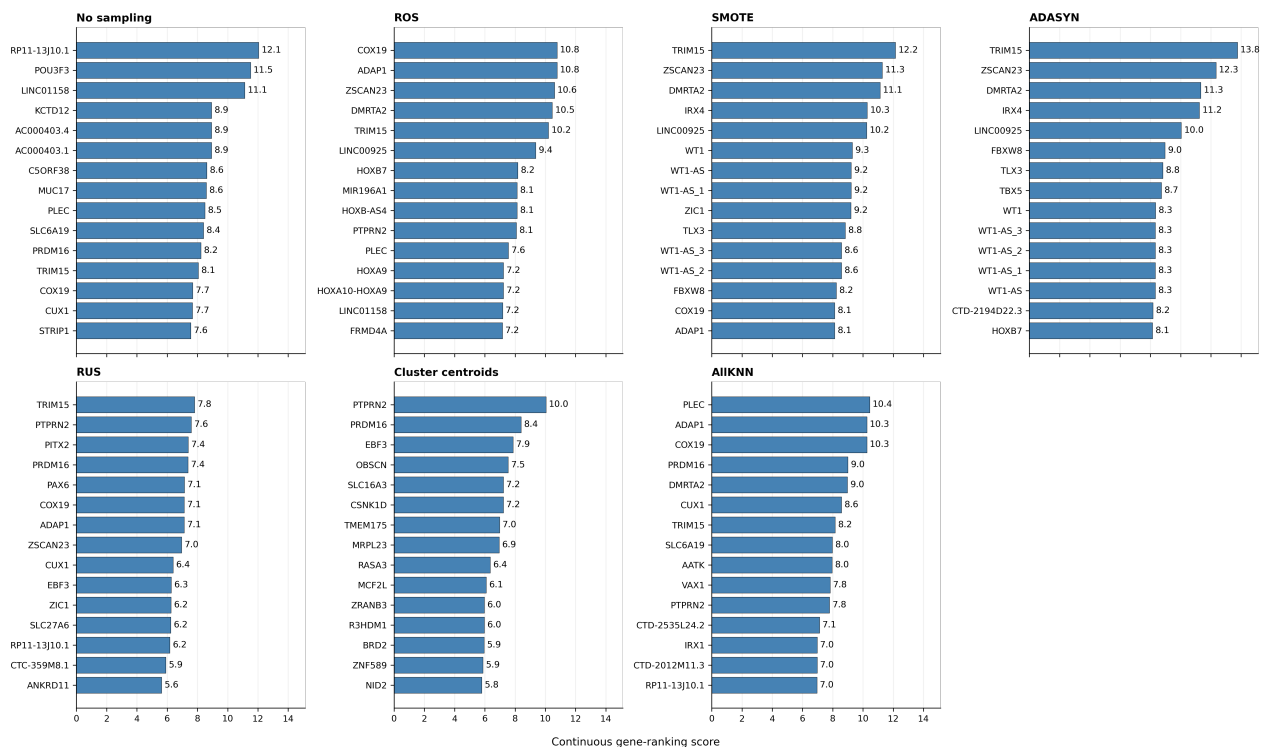


Figure S1. Highest-ranked genes identified under each class imbalance correction strategy. The top 15 genes according to the consensus continuous feature-ranking score are shown for each evaluated resampling strategy.

Table S1. Cross-method recurrence and method-specific highly ranked genes across class imbalance correction strategies. Genes shown appeared among the top 10 ranked genes under at least one experimental condition. Top-10 count indicates the number of the seven experimental conditions in which the gene appeared among the top 10. Top-10 methods identifies those conditions, and best rank indicates the highest rank achieved by the gene across the seven conditions.

Gene	Top-10 Count	Top-10 Methods	Best Rank
Recurrent top-10 genes			
TRIM15	5	ROS, SMOTE, ADASYN, RUS, AllKNN	1
ZSCAN23	4	ROS, SMOTE, ADASYN, RUS	2
DMRTA2	4	ROS, SMOTE, ADASYN, AllKNN	3
COX19	3	ROS, RUS, AllKNN	1
ADAP1	3	ROS, RUS, AllKNN	2
PRDM16	3	RUS, ClusterCentroids, AllKNN	2

Gene	Top-10 Count	Top-10 Methods	Best Rank
LINC00925	3	ROS, SMOTE, ADASYN	5
PLEC	2	No sampling, AllKNN	1
PTPRN2	2	RUS, ClusterCentroids	1
IRX4	2	SMOTE, ADASYN	4
WT1	2	SMOTE, ADASYN	6
CUX1	2	RUS, AllKNN	6
Method-specific top-10 genes			
RP11-13J10.1	1	No sampling	1
POU3F3	1	No sampling	2
LINC01158	1	No sampling	3
EBF3	1	ClusterCentroids	3
PITX2	1	RUS	3
AC000403.1	1	No sampling	4
OBSCN	1	ClusterCentroids	4
AC000403.4	1	No sampling	5
CSNK1D	1	ClusterCentroids	5
PAX6	1	RUS	5
FBXW8	1	ADASYN	6
KCTD12	1	No sampling	6
SLC16A3	1	ClusterCentroids	6
WT1-AS	1	SMOTE	7
TLX3	1	ADASYN	7
C5ORF38	1	No sampling	7
HOXB7	1	ROS	7
TMEM175	1	ClusterCentroids	7
WT1-AS_1	1	SMOTE	8
TBX5	1	ADASYN	8
MUC17	1	No sampling	8
HOXB-AS4	1	ROS	8
SLC6A19	1	AllKNN	8
MRPL23	1	ClusterCentroids	8
ZIC1	1	SMOTE	9
MIR196A1	1	ROS	9
AATK	1	AllKNN	9
RASA3	1	ClusterCentroids	9

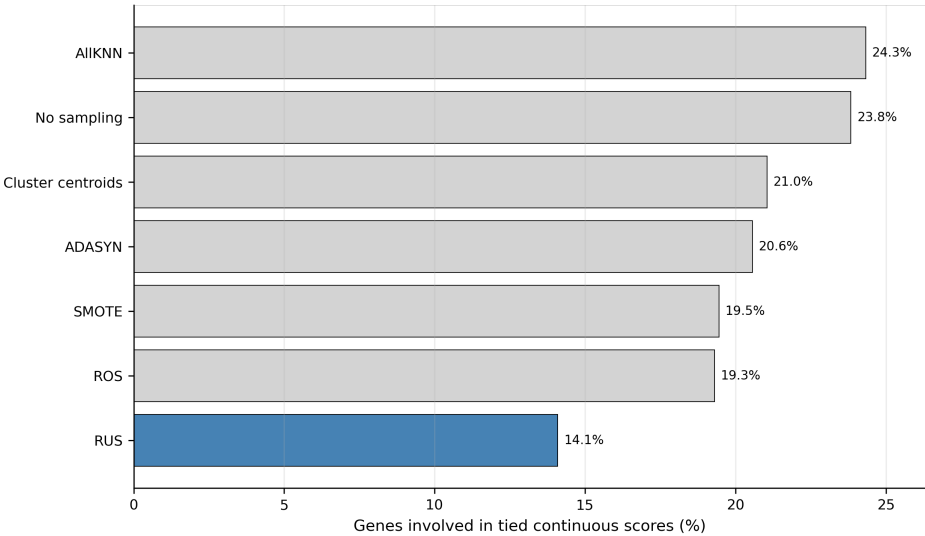


Figure S2. Feature-score tie diagnostics across resampling strategies. Bars show the percentage of genes involved in identical continuous consensus scores under each class imbalance correction strategy. Residual score ties ranged from 14.1% under RUS to 24.3% under AllKNN, illustrating method-dependent differences in ranking resolution.

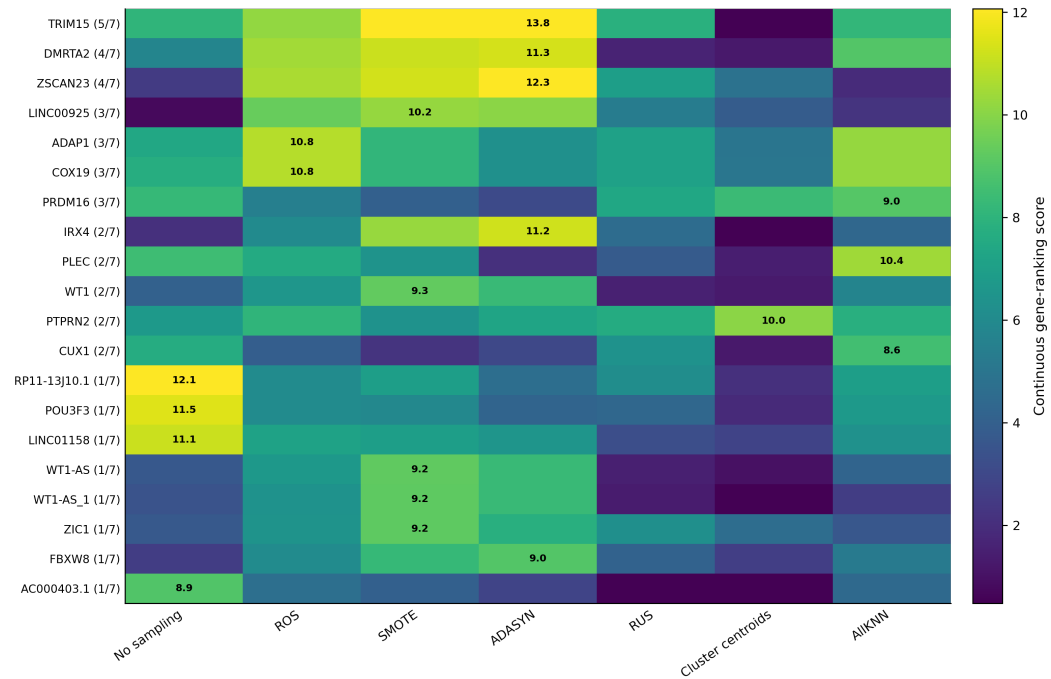


Figure S3. Cross-method recurrence of highly ranked genes across class imbalance correction strategies. Rows represent genes that appeared among the top 10 ranked genes for at least one experimental condition, and columns represent the no-resampling baseline and the six imbalance correction strategies. Heatmap colors indicate the continuous consensus score for each gene under each strategy, including strategies in which the gene did not rank among the top 10. Parenthetical values following each gene name indicate the number of the seven experimental conditions in which that gene appeared among the top 10. TRIM15 showed the greatest cross-method recurrence, appearing in the top 10 under five of seven conditions, followed by DMRTA2 and ZSCAN23 under four of seven conditions.

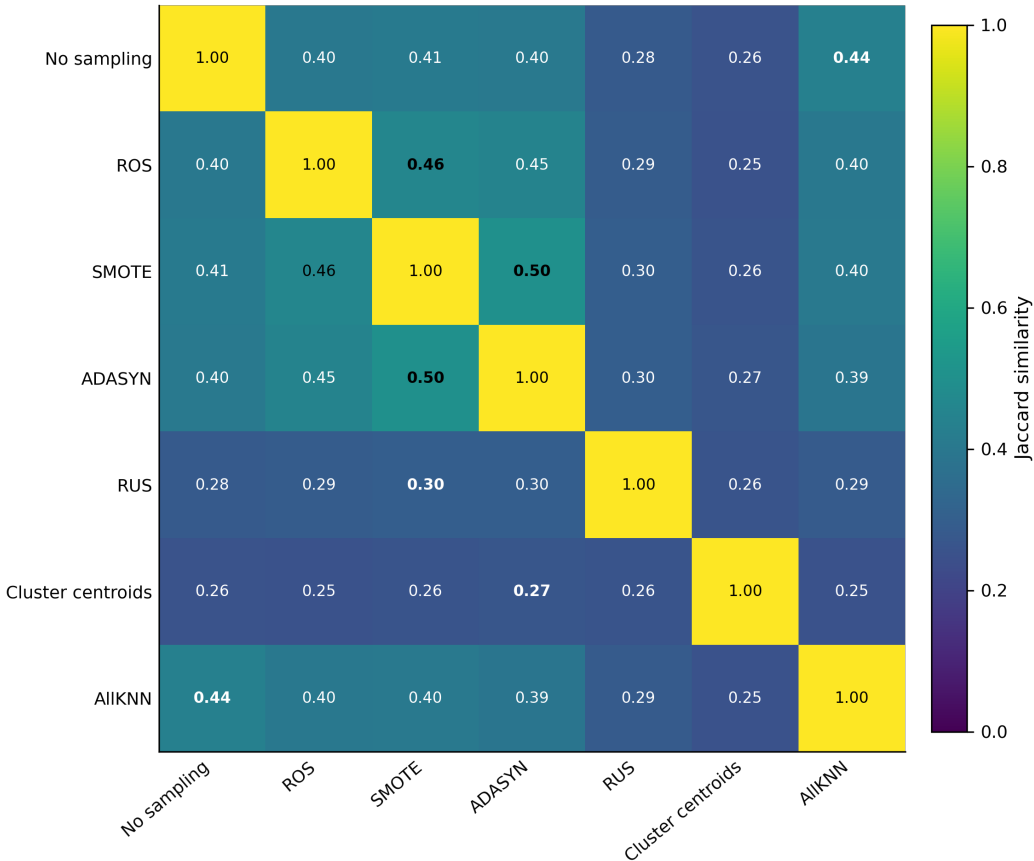


Figure S4. Pairwise similarity of consensus gene sets across class imbalance correction strategies. Pairwise Jaccard similarity coefficients quantify the overlap between consensus gene sets generated by each resampling strategy. Higher similarity values indicate greater agreement in prioritized biomarkers, whereas lower values highlight the influence of imbalance correction on feature selection.

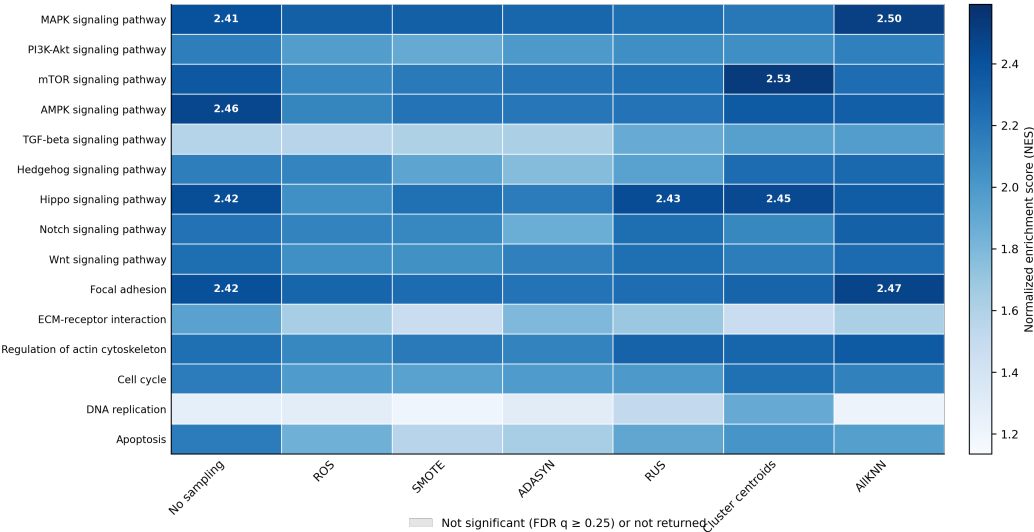


Figure S5. Additional KEGG pathway enrichment results across class imbalance correction strategies. Extended pathway enrichment analyses illustrating the consistency of biological mechanisms identified across the evaluated resampling strategies. The normalized enrichment scores support the robustness of the pathway-level conclusions presented in the main manuscript.