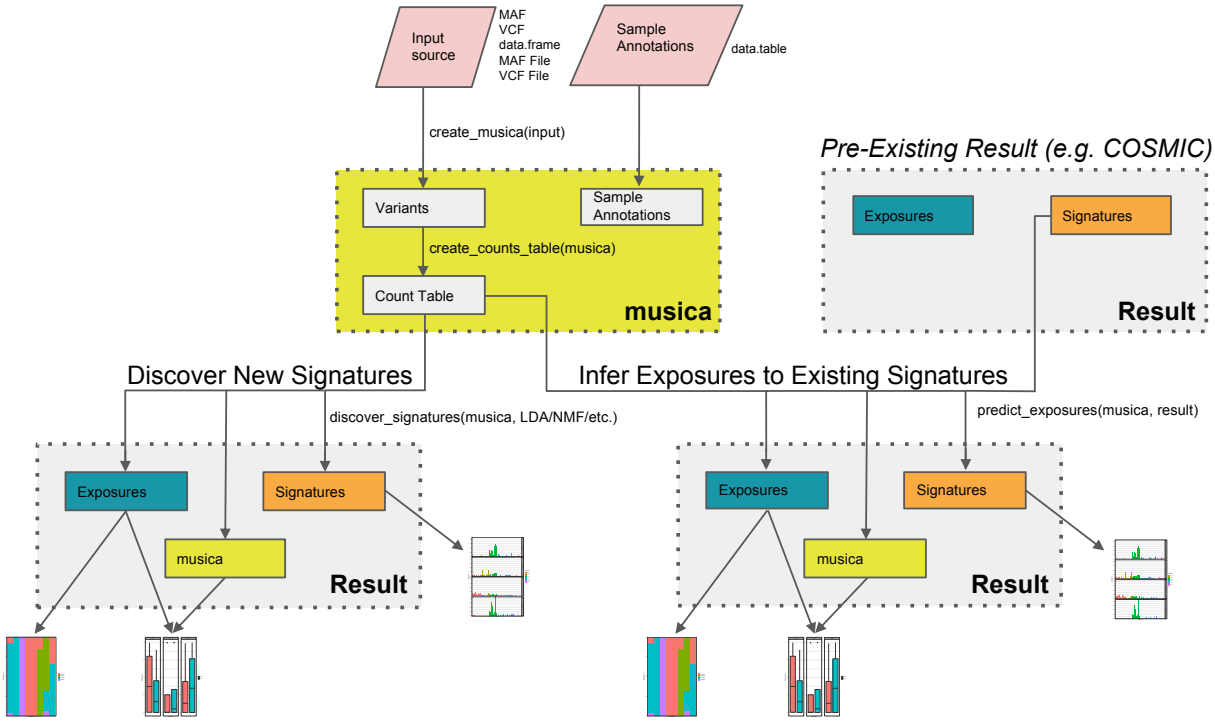
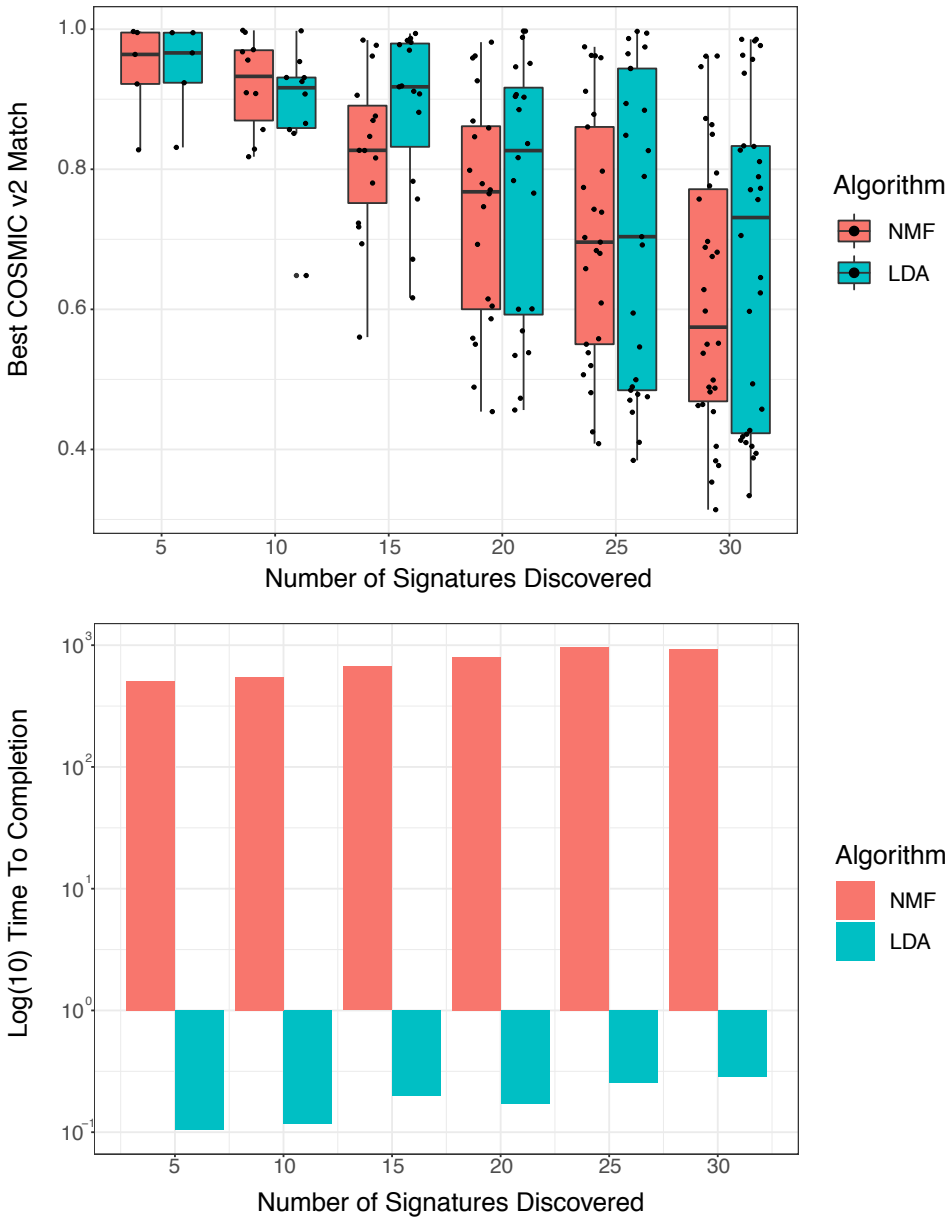
		musicatk	Signature Analyzer	SigProfiler	deconstruct Sigs	Signature Estimation	Mutational Patterns	YAPSA	decomp Tumor2Sig	Sigfit
Input	VCF	✓					✓		✓	
	MAF	✓								
	data.frame/table	✓								
	Mutation counts file	✓	✓	✓	✓	✓	✓	✓		✓
Inference	*de novo* discovery	✓	✓	✓			✓		✓	✓
	Prediction	✓			✓	✓	✓	✓	✓	✓
Motifs	snv96	✓	✓	✓	✓	✓	✓	✓	✓	✓
	snv192-transcript strand	✓		✓			✓			✓
	snv192-replication strand	✓					✓			
	dbs	✓	✓	✓			✓			
	indel	✓	✓	✓			✓	✓		✓
	Combine Features	✓	✓							
	Custom Features	✓								✓
Additional Features	Compare COSMIC v2	✓					✓	✓	✓	✓
	Compare COSMIC v3	✓					✓			✓
	Discovery by annotation	✓								
	Prediction by annotation	✓								
	Plot by annotation	✓						✓		
	UMAP	✓								
	Heatmap	✓					✓	✓		
	Clustering	✓					✓	✓		
	Differential Abundance	✓								

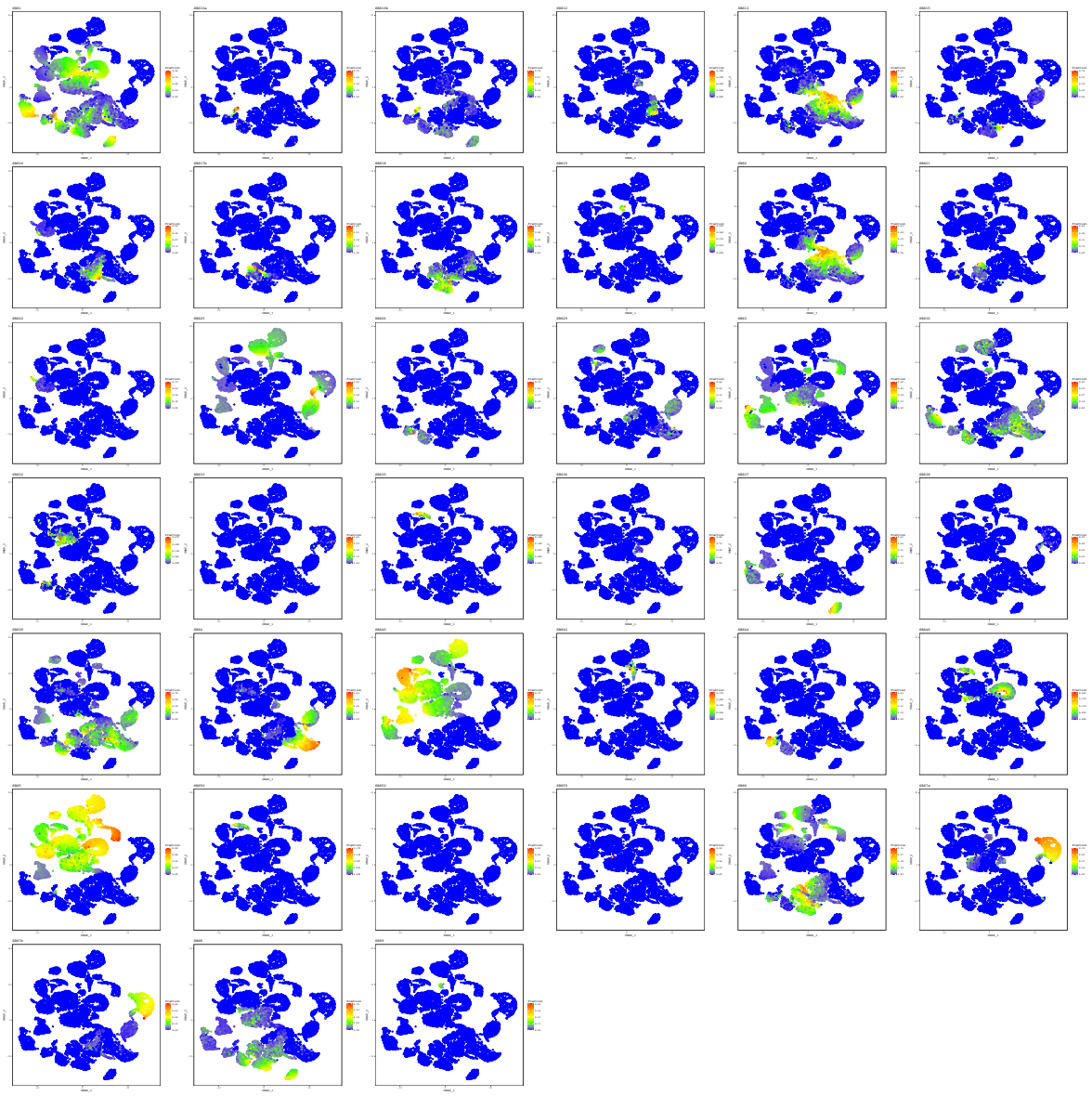
Supplementary Figure 1. Comparison of features across mutational signature packages.



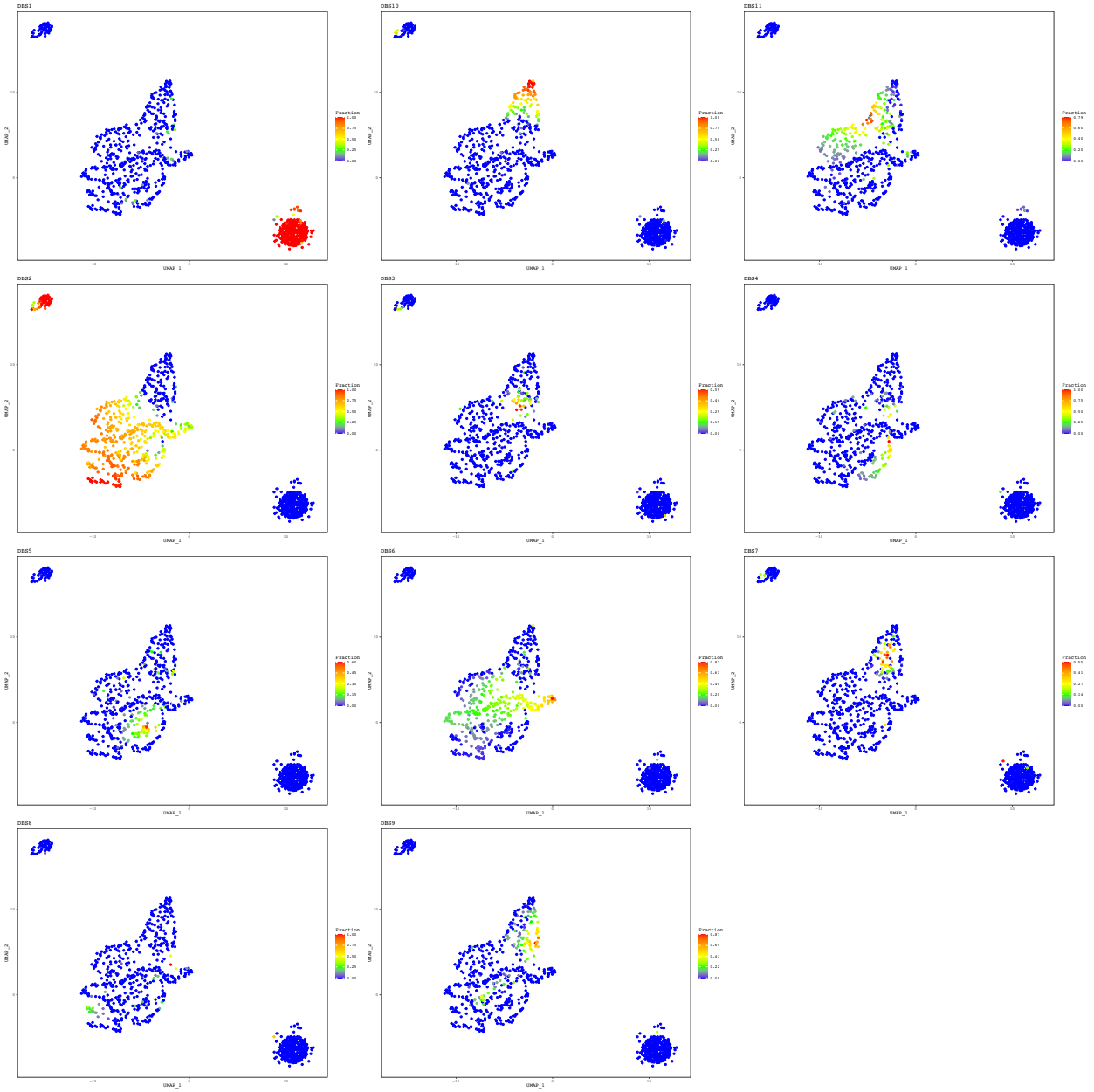
Supplementary Figure 2. Overview of the class structure in the *musicatk* package. This *musica* class is used as the main object for building count tables (SBS96, DBS78, IND83, SBS192, etc.) and for storing count tables for all variant classes. A *musica* object is used as input into the discovery or prediction functions. A *mustica_result* object is output from discover and prediction functions and used to store the variant tables along with the estimated signatures and exposures matrices. Existing signatures contained in a *mustica_result* object can be used to predict exposures in a new dataset. Result objects are available for COSMIC v2 and COSMIC v3 signatures in the package. User-generated *mustica_result* objects are used as input into all down-stream plotting and analysis functions.



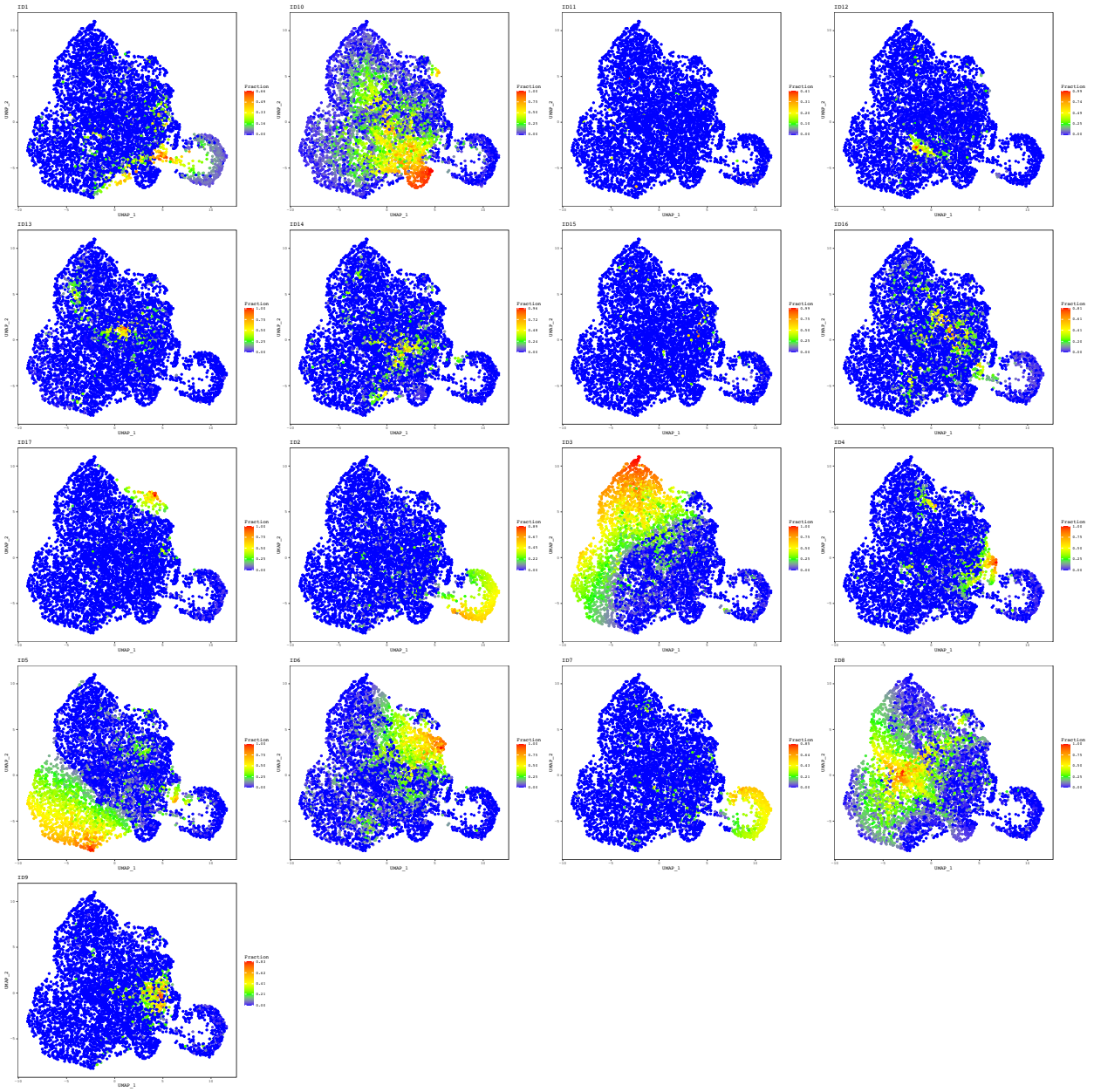
Supplementary Figure 3. Comparison of NMF and LDA to discover mutational signatures. We ran LDA and NMF to discover mutational signatures on the TCGA cohort with a range of k values (k = 5, 10, 15, 20, 25, and 30). **A)** Each signature was matched to the closest COSMIC v2 signature using cosine similarity. Boxplots show the distribution of cosine similarities from all signatures. A higher median cosine similarity was achieved with LDA for all runs except k=10, demonstrating that LDA can reconstruct signatures with accuracy similar to or better than NMF. **B)** The time in seconds required to perform the deconvolution for each algorithm shows that LDA substantially outperformed NMF in runtime.



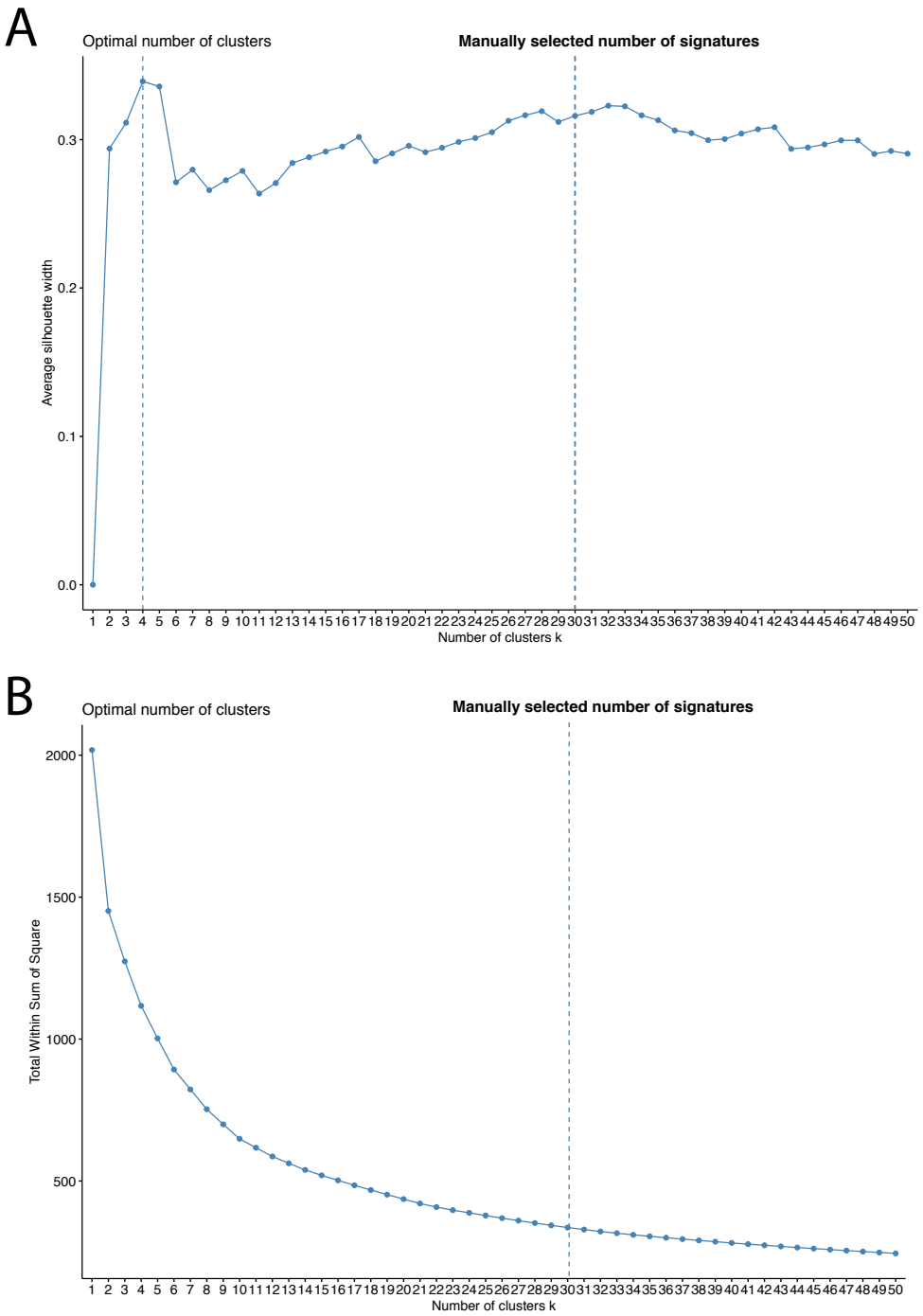
Supplementary Figure 4. UMAP colored by the predicted exposures from COSMIC v3 SBS signature. We applied the LDA-based prediction method to predict COSMIC v3 SBS signatures in a Pan-Cancer dataset from TCGA. 39 of the 65 signatures were found to be active in at least one tumor type. A UMAP plot was generated to explore the patterns of signatures across tumors Some signatures were present in nearly half of samples, some in a few tumor types, some in single tumor types, and some in subsets of multiple tumor types. For example, APOBEC signatures (SBS2 and SBS13) were present in a subset of tumors BRCA, CESC, BLCA, and HNSC, and distantly in a subset of LUAD and LUSC tumors. In general, both APOBEC signatures were present in the same samples. The only exceptions were disjoint subsets of BRCA samples with either SBS2 or SBS13. UCEC samples are split into three groups. The bottom samples are clustered separately because of their exposure to the SBS39 signature (unknown origin). The small top cluster includes samples from a few other tumor types including COAD and is defined by high levels of the POLE signature (SBS10 a/b). The left cluster has high levels of a defective DNA mismatch repair signature (SBS44). Two other MMR signatures, SBS15 and SBS21 could distinguish subsets of COAD and STAD tumors which also had higher levels of a third MMR signature, SBS6.



Supplementary Figure 5. UMAP colored by the predicted exposures from COSMIC v3 DBS signatures. All 11 of the COSMIC v3 DBS signatures were active in TCGA samples. Examining the DBS UMAP showed that DBS2 (tobacco smoke) is active in two of the 3 major clusters, representing ACC, LUAD, LUSC, HNSC, KIRP, LIHC, BLCA, ESCA, and MESO. DBS1 (UV light exposure) is found only in the SARC/SKCM cluster. DBS10 (defective DNA mismatch repair) is predominantly found in tumors from READ, PAAD, UCEC, and STAD and active in different sets of tumors from DBS1 and DBS2. DBS7 is also caused by defective DNA mismatch repair and mostly active in different sets of tumors from DBS10. The remaining signatures are present in mixed subsets of tumor types.



Supplementary Figure 6. UMAP colored by the predicted exposures from COSMIC v3 INDEL signatures. All 17 COSMIC v3 INDEL signatures were predicted to be active in TCGA samples. ID3 (tobacco smoking) was predominantly active in tumors from LUAD and LUSC. ID6 (defective DNA repair) was highly active in a distinct subset of samples containing mix of tumor types such as BRCA, OV, and STAD. High levels of ID10 (unknown etiology) defined a unique group of samples that were enriched for tumors from THCA and SARC. A distinct group of mixed tumor types was defined by different levels of activity for ID2 (defective DNA replication) and ID7 (defective DNA mismatch repair) suggesting that these aberrant processes may often co-occur.



Supplementary Figure 7. Metrics for selection of the optimal number of clusters. Two metrics provided to help users choose the optimal number of clusters. **A)** A higher average silhouette width indicates a better clustering solution. **B)** A lower "total within sum of squares" indicates a better clustering solution. The 4-cluster and 32-cluster solutions were the most and 2nd most stable solutions by silhouette, respectively. We choose the 30 cluster result as the final solution since the 32 cluster result splits the GBM-enriched (Glioblastoma Multiforme) cluster into two separate but similar clusters.