Supplementary Information

Biomarker discovery on synthetic patient transcriptomes: three public cohorts and a benchmark

Aditya Nanda

Supplementary Note 1. Borderline real DE genes and synthetic recall

Synthetic recall of real DE genes depends on how many real DE genes only just pass the real significance threshold. For each contrast we counted borderline real DE genes, those with real adjusted p from 0.01 to 0.05, and computed strict-rule recall separately for borderline genes and for the remaining real DE genes (Table S1). In ulcerative colitis versus control, 619 of 1,169 real DE genes (53.0%) were borderline, against 8.2% to 21.4% in the other anchored contrasts. Borderline genes were recovered less often than the remaining real DE genes in every contrast, and recall of real DE genes with adjusted $p < 0.01$ was 88.0–94.4% across anchored contrasts, including ulcerative colitis. Its lower overall recall therefore reflects its large share of borderline genes. STK11, not declared as an anchor column, had low recall even for real DE genes with adjusted $p < 0.01$ (60.7%).

Table S1. Borderline real DE genes and strict-rule recall in the full cohorts. Borderline, real DE genes with real adjusted p from 0.01 to 0.05. Recall is the percentage of real DE genes also called in the synthetic cohort at the strict rule (adjusted $p < 0.05$, $|\log_2 \mathrm{FC}| > 0.5$), shown for borderline genes, for real DE genes with real adjusted $p < 0.01$, and for all real DE genes (as in main-text Table 2). Non-anchored, a column used for DE but not declared as a dbTwin anchor.

	Real DE genes (n)		Recall, strict rule (%)		
Contrast	All	Borderline	Borderline	$p < 0.01$	All
TCGA-LUAD					
KRAS wild-type vs. mut.	1,121	159	65.4	88.0	84.8
EGFR wild-type vs. mut.	1,369	293	64.5	92.7	86.6
STK11 wild-type vs. mut. non-anchored	2,586	248	39.9	60.7	58.7
Sepsis					
Sepsis vs. control	2,743	224	60.7	93.0	90.4
RISK-IBD					
Crohn's disease vs. control	5,641	970	73.5	94.4	90.8
Ulcerative colitis vs. control	1,169	619	56.2	89.6	71.9

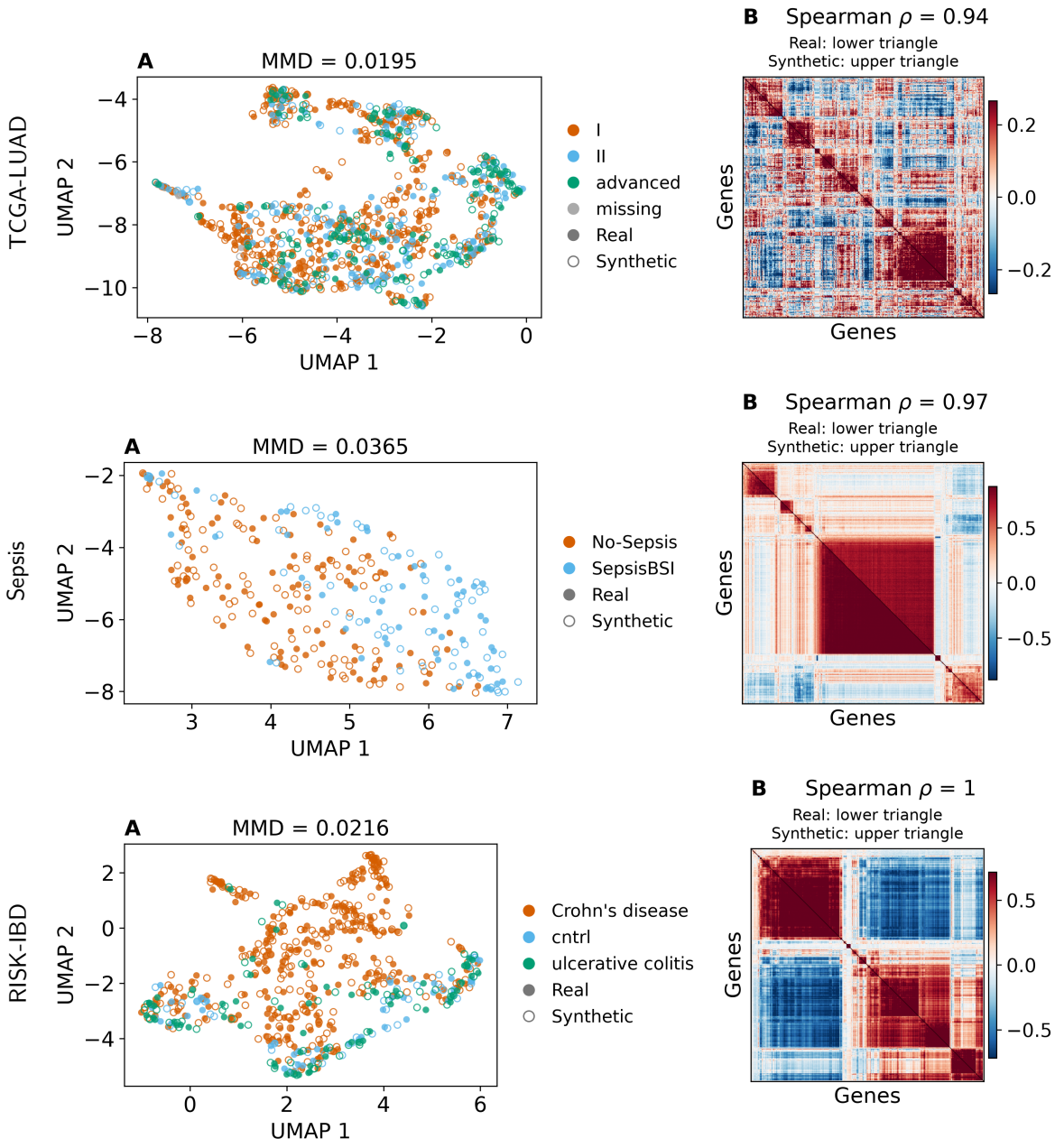


Figure S1. Expression fidelity in all three cohorts, as in the benchmark reports. Rows show TCGA-LUAD, Sepsis, and RISK-IBD. (A) Joint UMAP embedding of the full real (filled) and synthetic (open) cohorts, colored by ML-target class (gray, missing); MMD summarizes separation in the projected space. (B) Gene–gene correlation matrices for the 500 most variable real-cohort genes, with real correlations below and synthetic correlations above the diagonal; Spearman ρ summarizes off-diagonal agreement.

Supplementary Note 2. UMAP MMD and the number of embedding dimensions

The benchmark MMD is computed on a two-dimensional UMAP embedding, so it could depend on how much of the expression structure two dimensions retain. We therefore refit the joint UMAP of the full real and full synthetic cohorts with 3, 5, 7, 9, 11, 13, and 15 components, keeping the benchmark's $\log_2(\text{CPM} + 1)$ gene filter, 100 principal components, UMAP settings, and random state, and recomputed the RBF-kernel MMD (Fig. S2). The kernel bandwidth follows the dimension ($\gamma = 1/d$ for d components). From 3 to 15 dimensions, MMD ranged from 0.012 to 0.017 in TCGA-LUAD, 0.020 to 0.032 in sepsis, and 0.013 to 0.018 in RISK-IBD, against 0.020, 0.037, and 0.022 in two dimensions. In every cohort the two-dimensional benchmark gave the largest value, and MMD generally decreased as components were added.

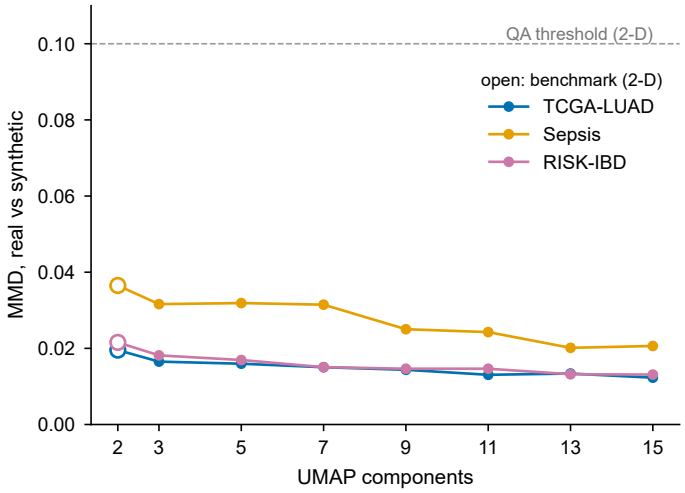


Figure S2. UMAP MMD against the number of embedding dimensions. RBF-kernel MMD between the full real and full synthetic cohorts in joint UMAP embeddings with 2 (the benchmark; open markers) to 15 components, all fitted to the same $\log_2(\text{CPM} + 1)$ genes and 100 principal components with the same UMAP settings and random state. Dashed line, the benchmark pass threshold of 0.1, which applies to the two-dimensional benchmark.

Supplementary Note 3. Holdout-based distance to closest record

The benchmark DCR uses no held-out data. As a complementary check, in every outer split we compared each synthetic record's distance to its nearest real training record with its distance to its nearest held-out real record (Fig. S3), using the benchmark preprocessing: $\log_2(\text{CPM} + 1)$ transformation, z-scoring with training statistics, and Euclidean distance over all genes. Because nearest-neighbor distances shrink as the reference set grows, the training set was subsampled to the holdout size (median over five subsamples). The ratio of fifth-percentile synthetic-to-training to synthetic-to-holdout distances was 0.897, 0.879, and 0.929 for TCGA-LUAD, sepsis, and RISK-IBD. In every cohort, synthetic records were therefore slightly closer to training records than to

held-out ones. This indicates a systematic pull toward the training data rather than near-copies: in 19,719 to 43,790 dimensions, nearest-neighbor distances are tightly concentrated, so a small shift is enough to move this ratio below one.

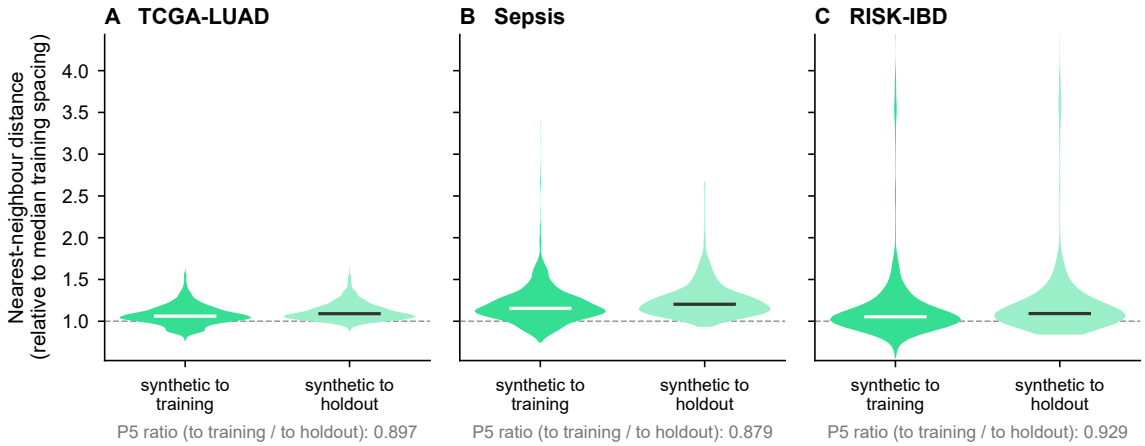


Figure S3. Holdout-based distance to closest record. For each cohort, distance from each synthetic record to its nearest real training record (green; training set subsampled to the holdout size, five subsamples pooled) and to its nearest held-out real record (light green), pooled over the five outer folds and scaled by each fold's median training-to-training distance (dashed line). Bars, medians; axis truncated at the 99.9th percentile. Below each panel, the fifth-percentile ratio of synthetic-to-training to synthetic-to-holdout distances, averaged over folds.