Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

ADITYA NANDA

aditya@dbtwin.ai

dbTwin https://orcid.org/0000-0002-6220-5250

Research Article

Keywords: transcriptomics, RNA-seq, clinico-transcriptomics, synthetic data generation, data-sharing, differential expression, machine learning, benchmark, patient stratification, biomarker discovery

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11732548/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare potential competing interests as follows: AN is a co-founder of and owns equity in dbTwin, Inc., which is developing the tool described in this manuscript.

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

Aditya Nanda [*1]

[1]dbTwin, Inc., USA

Abstract

Background: Patient-level clinico-transcriptomic cohorts are used for biomarker discovery, patient stratification, validation and model training. However, most transcriptomic data remains siloed in hospitals, biopharma, and consortia because expression profiles and linked clinical metadata carry re-identification risk. Synthetic RNA-seq cohorts that capture population structure without retaining individual records could be shared in their place, but only if they accurately reproduce the results of standard translational analyses on the real data. We test this with dbTwin, a proprietary generator, on three public cohorts, so that every synthetic result can be checked against its real source, and release the synthetic cohorts with the benchmark used to evaluate them.

Results: Synthetic versions of TCGA lung adenocarcinoma (510 samples; 19,719 genes), whole-blood sepsis (152; 27,097), and pediatric inflammatory bowel disease (322; 43,790) recovered most of the differential-expression (DE) signals tested. Across five contrasts on clinical variables declared to dbTwin as anchor columns, synthetic effect sizes kept their real magnitude ($\log_2$ fold-change slope 0.92–1.01), and 71.9–90.8% of real DE genes were recovered (81.7–97.2% at a relaxed screening threshold). For STK11 mutation (non-anchor column), effect sizes shrank (slope 0.70) but 96.8% of DE genes kept their direction. In five-fold cross-validation, classifiers trained on synthetic data were within 0.002 ROC-AUC of real-trained ones on held-out real patients and relied on largely the same driver genes (SHAP Pearson $r = 0.808$–0.975). Synthetic cohorts also retained gene–gene co-expression (Spearman $\rho = 0.940$–0.996), and 0 of 22 clinical comparisons exceeded sampling noise. Synthetic cohorts showed no sign of gross memorization by a distance-based check (fifth-percentile distance ratio 0.916–1.056), and attackers knowing a patient's age and sex gained little on training patients over unseen ones (excess success 0.028–0.066); those also knowing the patient's expression gained more, mainly on clinical labels (up to 0.518 for tumor stage), not on expression values (per-gene maximum 0.092–0.177).

Conclusions: On standard DE and machine-learning analyses, synthetic RNA-seq cohorts reproduced most real-data results on these three cohorts while retaining the joint structure of gene expression and clinical metadata. This suggests a possible route around data silos: a data custodian can generate and share a synthetic version of its cohort, and recipients can run their usual stratification and biomarker analyses on it and confirm candidates on the governed real data. The benchmark is generator-agnostic; its code and the real cohorts are publicly available at github.com/Nanda-Aditya/rna-twin, and the synthetic cohorts are freely and publicly available at huggingface.co/datasets/NandaAditya/rna-twin.

[*]Corresponding author: **aditya@dbtwin.ai**

Introduction

Patient-level clinico-transcriptomic cohorts are used across modern drug discovery for biomarker-discovery, patient-stratification, validation and model training. However, accessing external cohorts is slow and costly due to data-use agreements, institutional review, and controlled-access workflows [1, 2]. These controls exist for good reason: expression counts can retain genotype-associated information, while age, sex, tissue, disease status, and related fields act as quasi-identifiers that may strengthen linkage or inference attacks [3–6]. Trusted research environments and federated learning bring analyses to data that stay in place [7, 8], but heterogeneous technical stacks, governance, and export controls can limit interoperability and complicate analyses spanning organizations or platforms.

Synthetic cohorts offer a complementary route. A generated cohort that captures population structure without reproducing individual records can be shared for hypothesis generation, workflow development, and preliminary feature prioritization, with confirmatory analyses reserved for appropriately governed real data. For an analyst, such a cohort is useful only if standard analyses give the same answers as the real data: the same differentially expressed (DE) genes with the same effect sizes, classifiers that perform as well on real patients and rely on the same driver genes, and the same co-expression and clinical subgroup structure [9–11]. In other words, the synthetic cohorts must be biologically faithful - summary scores of distributional similarity cannot settle these questions on their own, because a synthetic cohort can match marginal distributions and broad correlation structure while still distorting the specific signals a scientist cares about. Privacy evaluation must likewise extend beyond a single proximity statistic [12–14].

dbTwin (dbTwin, Inc.) is a proprietary ML-based generator for clinico-transcriptomic data that operates directly on full-transcriptome expression and clinical metadata tables. Users can declare the clinical variables they intend to analyze as anchor columns, and dbTwin prioritizes their relationships with expression. We previously introduced dbTwin together with a transcriptome-scale benchmark that compared a beta version of the generator with statistical and deep-generative baselines on DE fidelity, ML performance, gene-level feature attribution, and distance-based memorization [11]. That benchmark tested DE and ML on a single contrast column and measured privacy with a single distance-to-closest-record (DCR) statistic. Here we evaluate the updated generator with a broader benchmark that adds clinical metadata fidelity, six DE contrasts across three cohorts, a separate prediction task per cohort, and attribute-inference attacks.

We applied dbTwin to three public cohorts spanning lung adenocarcinoma, sepsis, and pediatric inflammatory bowel disease. We chose public cohorts deliberately: anyone can check every synthetic result against its real counterpart, and the same workflow can then be applied to cohorts that cannot be shared. We evaluate each synthetic cohort with the analyses a translational analyst would run on the real data, DE contrasts and a prediction task, and assess privacy with distance-based and attribute-inference analyses (Fig. 1). We release the three synthetic cohorts, each as a full cohort for use and as five fold-specific cohorts for the held-out analyses, together with a generator-agnostic benchmark that grades a synthetic cohort on these analyses and a benchmark report for each cohort.

The benchmark code and real cohorts are available at github.com/Nanda-Aditya/rna-twin and the synthetic cohorts at huggingface.co/datasets/NandaAditya/rna-twin.

Results

Cohorts and benchmark overview

We generated synthetic versions of three public clinico-transcriptomic cohorts spanning lung cancer, sepsis, and pediatric inflammatory bowel disease (Table 1): 510 TCGA-LUAD tumor samples measured over 19,719 genes, 152 whole-blood samples and 27,097 genes from the GSE189400 (EARLI) sepsis cohort, and 322 ileal biopsy samples and 43,790 genes from the pediatric RISK cohort (RISK-IBD) [15–19]. Users can declare up to eight translationally meaningful clinical variables as anchor columns, on which dbTwin prioritizes fidelity; the anchor columns used for each cohort, covering mutation status, disease labels, age, and sex, are listed in Table 1.

Each cohort was synthesized in two ways (Fig. 1). The full synthetic cohort, generated from all real samples, is the released dataset; differential expression, global expression structure, and clinical fidelity were evaluated on it against the full real cohort. Classifiers and privacy analyses need held-out real patients that the generator has never seen, so for these we split each real cohort into five stratified folds and generated one synthetic cohort per fold from the training samples alone, keeping the held-out real fold for evaluation. Both the full and the fold-specific synthetic cohorts are publicly released.

The benchmark is generator-agnostic: it needs only the real cohort, the synthetic cohorts, and a configuration file naming the DE contrasts, prediction task, anchor columns, and quasi-identifiers, so it can evaluate synthetic data from any generator. For each cohort it produces a report that grades the synthetic data against fixed pass thresholds: gene–gene co-expression Spearman $\rho \geq 0.8$ and UMAP maximum mean discrepancy (MMD) ≤ 0.1; for each anchored DE contrast, $\log_2$ fold-change Pearson $r \geq 0.8$ and recall of at least 80% of real DE genes at the relaxed rule (below); and, for the prediction task, a synthetic-trained classifier within 0.05 ROC-AUC of a real-trained one on held-out real patients, with SHAP Pearson $r \geq 0.8$. Analyses on columns not declared as anchor columns are reported but not graded, and privacy metrics are reported without a pass threshold.

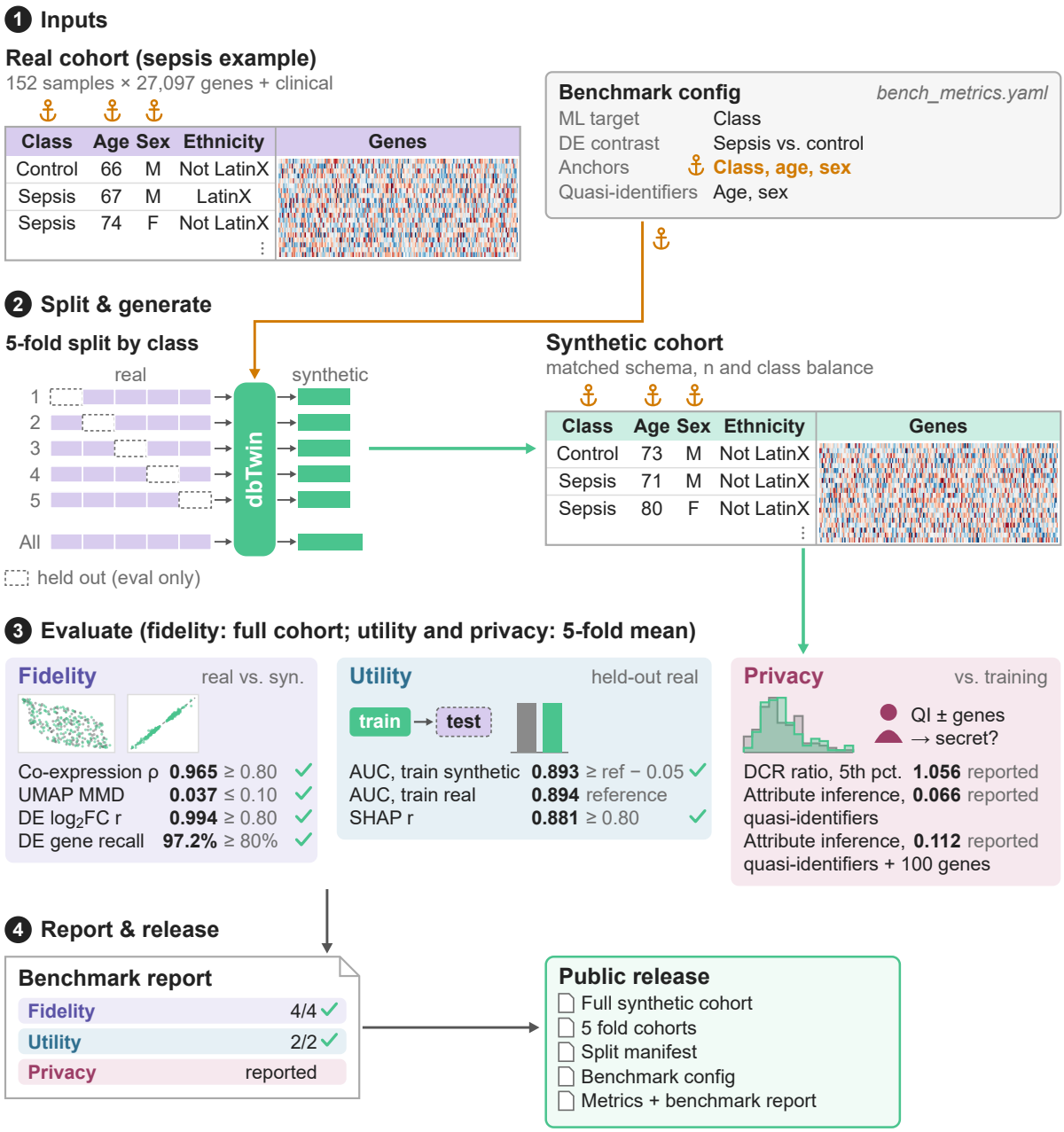


Figure 1. Synthetic cohort generation and validation. Each source cohort is synthesized in two ways. For the released full cohort, dbTwin is fit on all real samples, and differential-expression and global fidelity are evaluated on it. For the prediction task and privacy analyses, each source cohort is divided into five stratified outer folds; for each fold, dbTwin generates a synthetic cohort with matched schema, sample size, and class balance from the training samples, while the held-out real fold is locked for evaluating utility and privacy. In the sepsis example shown, the prediction task (ML target) is Class, that is, sepsis status. dbTwin also accepts optional anchor columns (marked with an anchor icon): clinical variables whose relationships with gene expression are preserved with high fidelity. Real and synthetic data are compared for global statistical fidelity (co-expression and UMAP embedding), differential-expression fidelity (DE gene overlap and fold-change correlation), machine-learning fidelity (AUC of classifiers trained on synthetic and tested on real data, TSTR, and SHAP concordance), and privacy (DCR and attribute inference), and the results are compiled into figures and a benchmark report for each cohort.

Table 1. Source cohorts, anchor columns, and evaluation tasks. Counts refer to the complete source cohorts. CD, Crohn's disease; UC, ulcerative colitis.

	TCGA-LUAD	**Sepsis**	**RISK-IBD**
Source	TCGA-LUAD	GSE189400	GSE57945 (RISK)
Tissue	Lung tumor	Whole blood	Ileal biopsy
Samples	510	152	322
Genes	19,719	27,097	43,790
Anchor columns	EGFR mutation, KRAS mutation, age, sex, tumor stage	Sepsis status, age, sex	Disease class, age at diagnosis, sex
DE contrasts	KRAS, EGFR, and STK11[†]: wild-type vs. mutant	Sepsis vs. control	CD vs. control; UC vs. control
Prediction task	Tumor stage (I, II, advanced)	Sepsis vs. control	Disease class (CD, UC, control)

[†]STK11 mutation is not an anchor column.

Synthetic cohorts recover real differential-expression signals

Differential expression is usually the first analysis run on a new clinico-transcriptomic cohort, which makes it the most direct test of whether a synthetic cohort returns the same discoveries. Of the six prespecified contrasts, five were defined on anchor columns (anchored contrasts) and one, STK11 mutation, on a column that was not. For each, we ran DESeq2 independently on the full real and synthetic cohorts and asked two questions: are real effect sizes reproduced, and are real DE genes recovered (Table 2; Fig. 2)? Real DE genes required adjusted $p < 0.05$ and $|\log_2 \mathrm{FC}| > 0.5$. Synthetic DE genes were called with the same strict rule, or with a relaxed rule (adjusted $p < 0.075$, $|\log_2 \mathrm{FC}| > 0.4$) intended for exploratory screening, where missing a real signal costs more than following a false lead.

Effect sizes were reproduced in all five anchored contrasts. Over the real DE genes, synthetic and real $\log_2$ fold changes correlated at $r = 0.961$–0.994, and the slope of synthetic on real fold changes was 0.92–1.01 (Fig. 2C). In Crohn's disease versus control, for example, the 5,641 real DE genes kept fold changes close to their real values ($r = 0.993$, slope 1.00; Fig. 2A). Slopes near one indicate that real effects were kept close to their real magnitude rather than shrunk toward zero.

Gene recovery varied more than effect sizes. At the strict rule, synthetic cohorts recovered 71.9–90.8% of real DE genes with precision of 75.1–89.9% (Fig. 2D): most synthetic DE genes were real DE genes, and confirmation on real data would screen out the rest. Recovery was highest for sepsis (90.4%) and Crohn's disease (90.8%) and lower for KRAS mutation (84.8%), EGFR mutation (86.6%), and ulcerative colitis (71.9%). Group size alone did not set this order: sepsis compared 60 with 92 subjects yet recovered a larger share of its real DE genes than KRAS, with 154 mutant tumors. Recovery was instead lowest where many real DE genes only just passed the real threshold. Real DE genes with adjusted $p < 0.01$ were recovered at 88.0–94.4% in every anchored contrast, whereas borderline genes, with real adjusted p from 0.01 to 0.05, were recovered less often. Ulcerative colitis, with the lowest recall, had the largest share of borderline genes

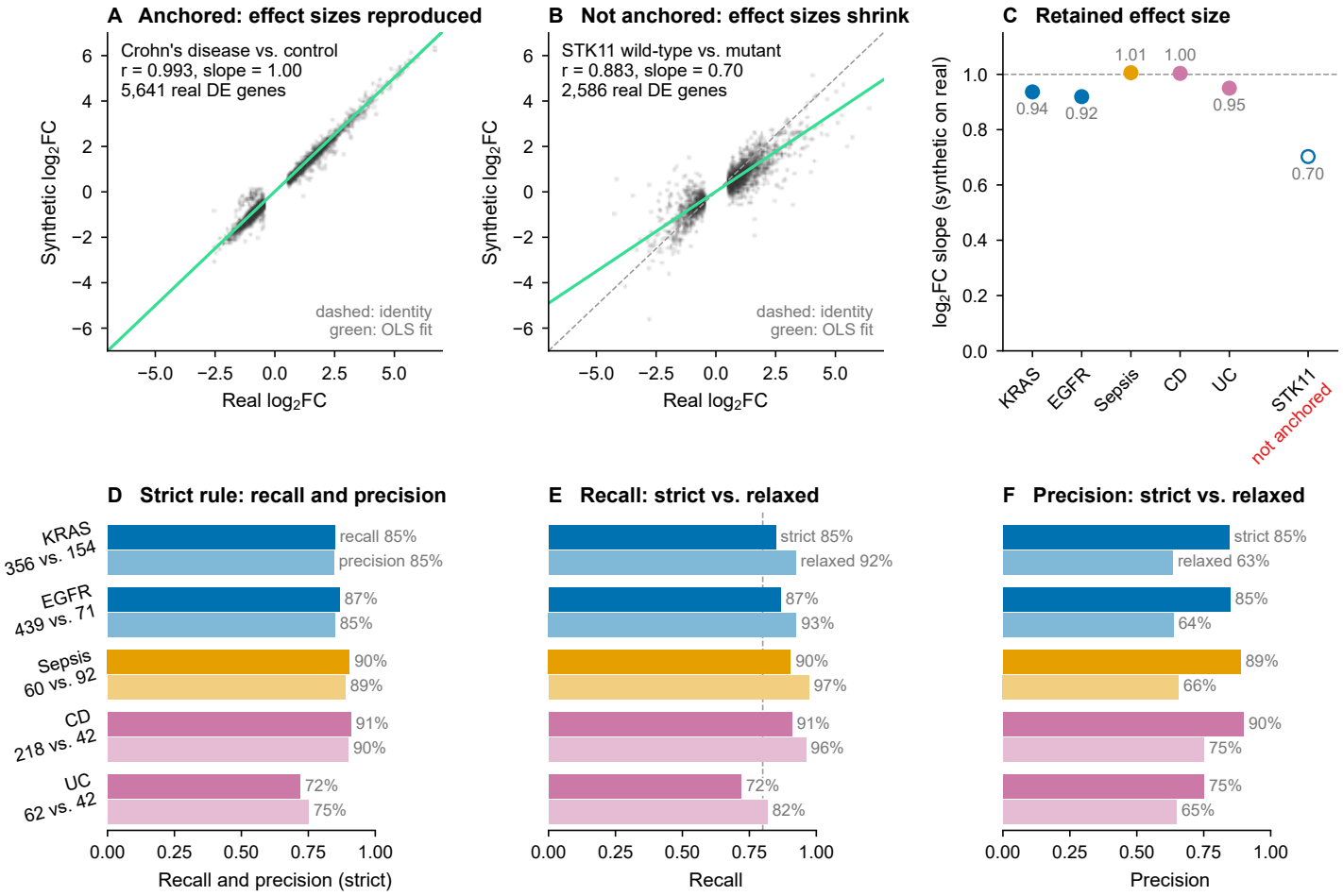


Figure 2. Differential-expression fidelity Colors denote cohorts (blue, TCGA-LUAD; yellow, sepsis; pink, RISK-IBD). (A) Real versus synthetic $\log_2$ fold changes for the 5,641 real DE genes of Crohn's disease versus control, an anchored contrast. Dashed line, identity; green line, least-squares fit. (B) As in A for STK11 wild-type versus mutant in TCGA-LUAD (2,586 real DE genes), not declared as an anchor column; the fit is shallower than the identity line. (C) Slope of synthetic on real $\log_2$ fold changes over real DE genes for every contrast. Filled circles, anchored contrasts; open circle, not anchored; dashed line, slope of one (effect sizes fully retained). (D) Recall (dark bars) and precision (light bars) of synthetic DE genes at the strict rule for the five anchored contrasts. Labels give group sizes (case vs. reference), ordered from the largest to the smallest minority group. (E) Recall at the strict (dark) and relaxed (light) synthetic rules; dashed line, benchmark pass threshold of 0.8 for relaxed-rule recall. (F) Precision at the strict (dark) and relaxed (light) rules. Rules as in Table 2.

(53.0%, against 8.2% to 21.4% in the other anchored contrasts; Supplementary Note 1). Because effect sizes were preserved, most missed genes are likely real effects that fell just short of synthetic significance. Consistent with this, the relaxed rule raised recovery to 81.7–97.2% (Fig. 2E), at the cost of precision, which fell to 63.4–75.1% (Fig. 2F). We therefore recommend the relaxed rule for screening, with candidates confirmed on real data.

Table 2. Differential-expression fidelity in the full real and synthetic cohorts. Real DE genes: adjusted $p < 0.05$ and $|\log_2 \mathrm{FC}| > 0.5$. Synthetic DE genes: strict, same rule; relaxed, adjusted $p < 0.075$ and $|\log_2 \mathrm{FC}| > 0.4$. Recall is the percentage of real DE genes also called in the synthetic cohort; precision is the percentage of synthetic DE genes that are real DE genes. Non-anchored, a contrast on a column not declared as an anchor column. $\mathrm{Log_2FC}$ r and slope are, respectively, Pearson correlation and least-squares slope of synthetic on real $\log_2$ fold changes over all real DE genes.

					Recall (%)		Precision (%)	
Contrast	n (case / ref.)	Real DE	$\mathrm{log_2FC}$ r	Slope	strict	relaxed	strict	relaxed
TCGA-LUAD								
KRAS wild-type vs. mut.	356 / 154	1,121	0.961	0.94	84.8	92.3	84.5	63.4
EGFR wild-type vs. mut.	439 / 71	1,369	0.970	0.92	86.6	92.5	84.9	63.7
STK11 wild-type vs. mut. non-anchored	435 / 75	2,586	0.883	0.70	58.7	69.1	81.1	72.3
Sepsis								
Sepsis vs. control	60 / 92	2,743	0.994	1.01	90.4	97.2	88.7	65.6
RISK-IBD								
Crohn's disease vs. control	218 / 42	5,641	0.993	1.00	90.8	96.1	89.9	75.1
Ulcerative colitis vs. control	62 / 42	1,169	0.980	0.95	71.9	81.7	75.1	64.8

STK11 mutation was the only contrast defined on a column that was not provided to dbTwin as an anchor column. Its synthetic effect sizes were shrunk (slope 0.70, $r = 0.883$; Fig. 2B), and only 58.7% of its 2,586 real DE genes were recovered at the strict rule (69.1% at the relaxed rule). The direction of regulation, however, largely stayed intact: 96.8% of real DE genes changed in the same direction in the synthetic cohort. This matches the intended role of anchor columns: relationships with anchor columns are prioritized, whereas associations with other columns are preserved only partially.

Synthetic-trained classifiers match real-trained ones on unseen patients

We next asked whether models trained on synthetic data generalize to real patients they have never seen. For each outer fold, we trained elastic-net classifiers on the fold-specific synthetic cohort and evaluated them on the corresponding held-out real fold (train on synthetic, test on real; TSTR), alongside real-trained models evaluated on the same patients (TRTR). The inverse regularization strength C was chosen by cross-validation on real training data alone, as the value with the highest AUC among those that kept 20–60% of input genes active (Methods).

TSTR closely matched TRTR in all three cohorts, with ROC-AUC differing by at most 0.002 (Fig. 3A). Tumor stage in TCGA-LUAD is only weakly predictable from bulk expression even with

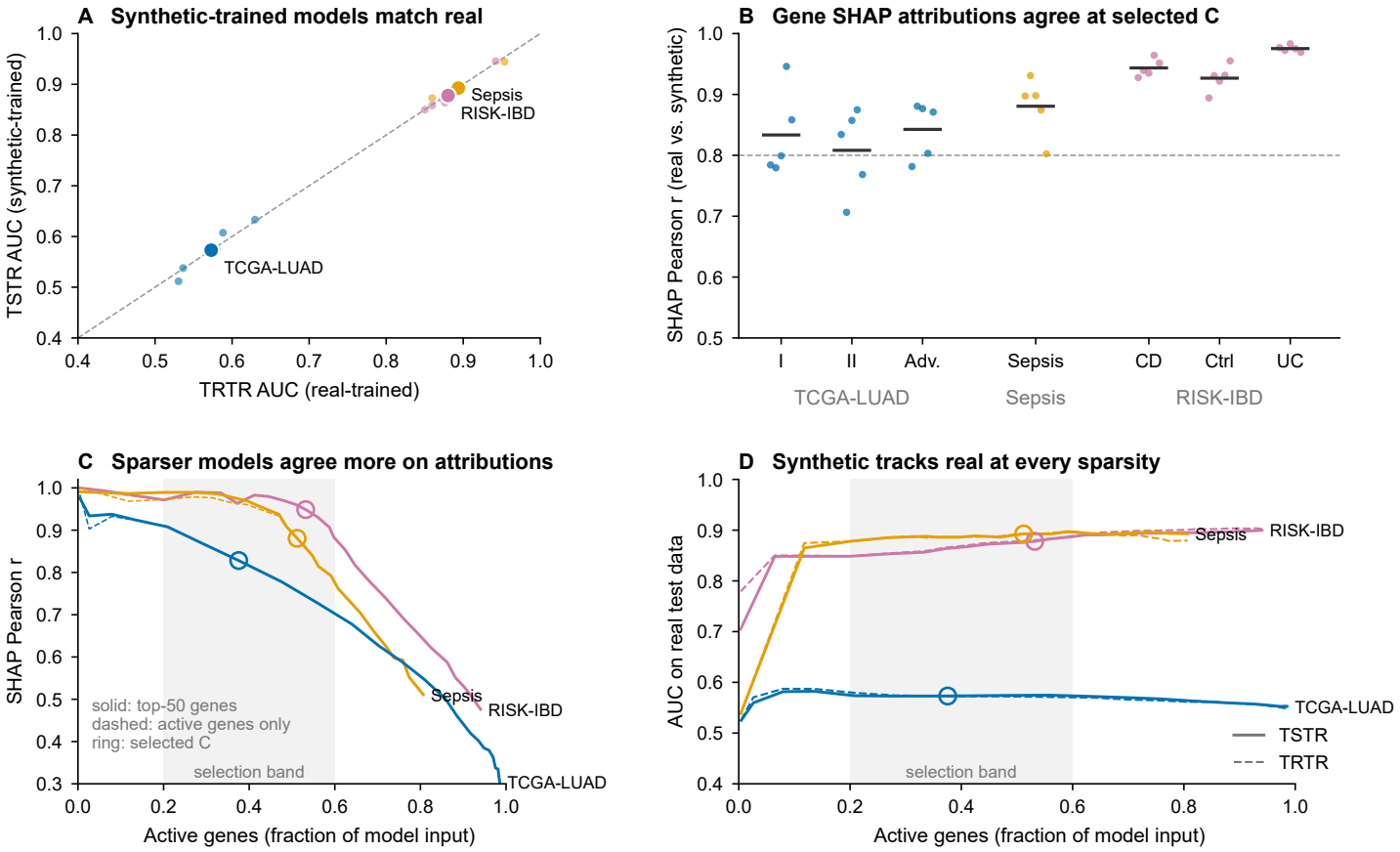


Figure 3. Machine-learning fidelity on held-out real patients. Colors denote cohorts as in Fig. 2. (A) ROC-AUC of synthetic-trained (TSTR) versus real-trained (TRTR) elastic-net classifiers, both evaluated on the same held-out real fold. Small points, outer folds; large points, cohort means; dashed line, identity. (B) SHAP concordance per class at the selected C: Pearson r between mean absolute SHAP values of real- and synthetic-trained models over the 50 most important real-trained genes, computed on held-out real samples. Points, outer folds; bars, means; dashed line, 0.8. (C) SHAP concordance (mean over folds and classes) against model sparsity, from refitting both models (real-trained and synthetic-trained) over 28 values of C at a fixed ℓ_1 ratio of 0.05. The x-axis is the mean fraction of input genes with a non-zero real-trained coefficient. Solid lines, SHAP correlation with top 50 real-trained genes; dashed lines, top $\min(50, n_{\text{active}})$ genes, so that inactive genes with zero attribution cannot inflate r (the two differ only for the sparsest models). Rings, selected C; gray band, 20–60% active-gene band, within which C is selected as the value with the highest inner cross-validated AUC on real training data. (D) TSTR (solid) and TRTR (dashed) ROC-AUC on held-out real folds across the same sweep.

real data (TRTR AUC 0.573); dbTwin reproduced this ceiling rather than inflating it (TSTR AUC 0.573).

Matching AUC is not enough, because two classifiers can reach the same AUC using different genes. We therefore compared the genes each model relied on, using SHAP attributions computed on held-out real samples. The leading genes largely agreed: on average, 72.0%, 74.0%, and 84.0% of the ten most important real-trained genes were also among the ten most important synthetic-trained genes in TCGA-LUAD, sepsis, and RISK-IBD. Over the 50 most important real-trained genes, attributions were concordant in every class of every cohort (SHAP Pearson $r = 0.808$–0.975; Fig. 3B). Because SHAP values are computed on real patients excluded from generation, this concordance indicates that synthetic-trained models relied on largely the same genes as real-trained models when applied to unseen patients.

SHAP concordance and AUC both depend on how many genes a model uses, which is set by C. Refitting real- and synthetic-trained models over 28 values of C showed that TSTR tracked TRTR across the whole range, so agreement in predictive performance does not depend on this choice, whereas SHAP concordance was highest for sparse models and declined as more genes entered the model (Fig. 3C,D; Methods).

Synthetic cohorts preserve global expression and clinical structure

The DE and classifier results rest on the synthetic cohorts reproducing the basic structure of the data, so we also ran the checks an analyst would apply to any new cohort (Fig. 4). Per-gene means ($r = 1.000$, 1.000, and 1.000 in TCGA-LUAD, sepsis, and RISK-IBD) and the mean-variance relationship of count data (median synthetic-to-real variance ratio 0.98, 1.08, and 1.06) both matched the real cohorts (Fig. 4A,B), so the synthetic counts were not smoothed toward a simpler distribution. Gene–gene correlations among the 500 most variable genes were retained (Spearman $\rho = 0.940$–0.996; Fig. 4C), and real and synthetic samples intermixed in a joint UMAP embedding (MMD = 0.020–0.037; Fig. 4D; all cohorts in Supplementary Fig. S1); MMD was no larger when the embedding had 3 to 15 dimensions instead of two (Supplementary Note 2). Known sex differences were also reproduced. For a panel of sex-chromosome genes, including XIST and the Y-linked RPS4Y1, DDX3Y, and KDM5D, the slope of the synthetic on the real male-minus-female difference in mean $\log_2(\text{CPM} + 1)$ was 0.99, 0.99, and 1.00 in TCGA-LUAD, sepsis, and RISK-IBD. Sex is an anchor column in all three cohorts, so this shows that the known sex-linked signal of a declared anchor was carried into the synthetic data.

At the clinical level, we compared the distributions of the other anchor columns within every prediction class by total variation distance (TVD; 0 for identical distributions, 1 for disjoint ones). Because even a perfect generator yields a non-zero TVD in finite samples, we judged each value against the TVD between the real class and same-size resamples of itself. Categorical anchor columns were nearly identical (maximum TVD per cohort 0.000–0.015), whereas binned age differed more (0.048–0.214). Judged against sampling noise, 0 of the 22 class-by-anchor comparisons exceeded the 95% resampling range (Fig. 4E).

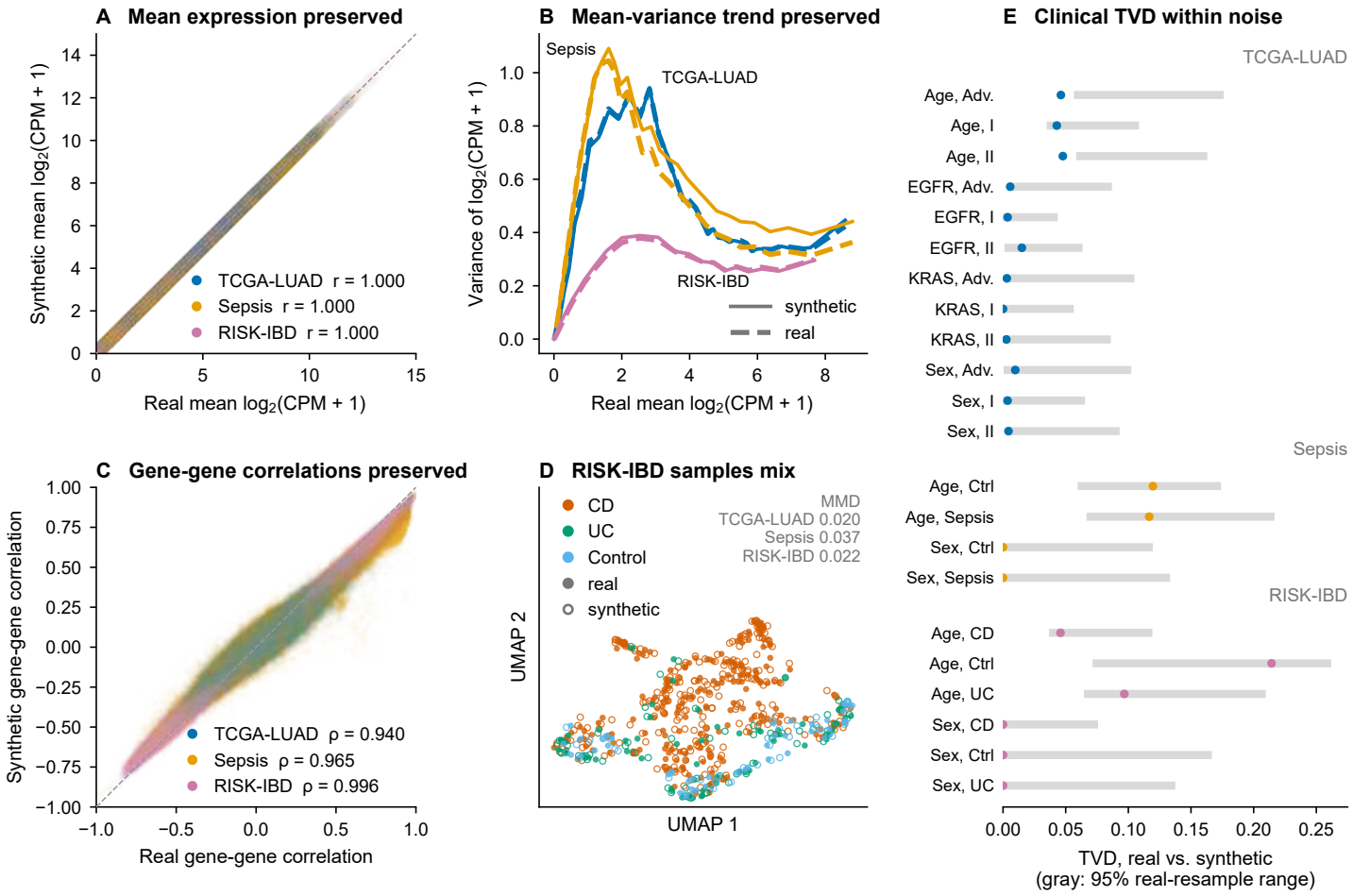


Figure 4. Statistical and clinical fidelity of the full synthetic cohorts. Colors in A to C and E denote cohorts (blue, TCGA-LUAD; yellow, sepsis; pink, RISK-IBD). (A) Mean $\log_2(\text{CPM} + 1)$ of every gene in the real (x) and synthetic (y) cohort; dashed line, identity; key, Pearson r. (B) Mean-variance relationship: genes are grouped into 40 equal-count bins by real mean expression, and lines give the median per-gene variance of $\log_2(\text{CPM} + 1)$ in each bin for real (dashed) and synthetic (solid) data. (C) Real versus synthetic Pearson correlation for every pair of the 500 most variable genes, one semi-transparent point per pair, so denser regions appear darker; dashed line, identity; key, Spearman ρ over all pairs. (D) Joint UMAP embedding of real (filled) and synthetic (open) RISK-IBD samples, colored by disease class; MMD for all three cohorts in the corner (other cohorts in Supplementary Fig. S1). (E) Total variation distance between real and synthetic distributions of each anchor column other than the class label, within each prediction class (dots). Gray bars give the 95% range of the same distance between the real class and 500 same-size bootstrap resamples of itself, the difference expected from sampling alone.

Synthetic cohorts show no sign of gross memorization, and attribute-inference gains concentrate in clinical labels

Finally, we evaluated two distinct privacy risks on the fold-specific cohorts: whether the generator memorized its training records, and whether a synthetic cohort helps an attacker infer something new about the real patients used to build it (Fig. 5).

We first used distance to closest record (DCR) as a memorization diagnostic. A generator that copies or lightly perturbs training records produces synthetic samples that sit unusually close to them, so we compared each synthetic sample's distance to its nearest training record with the ordinary nearest-neighbor spacing among training records themselves (Methods); ratios near one indicate no excess of near-copies. Fifth-percentile ratios were 0.916–1.056 and median ratios 0.949–1.007 (Fig. 5A). Synthetic records were, however, slightly closer to training than to held-out records (holdout ratio 0.879 to 0.929; Supplementary Note 3), a systematic pull toward the training data with no evidence of gross memorization. A ratio near one does not by itself establish privacy, however; it only rules out the most direct failure mode.

Attribute inference asks a more operational question. An attacker holds the released synthetic cohort, knows some facts about a real patient in the training data, and tries to infer something they do not know (a secret): another clinical field, or the expression of the next 100 most variable genes. We considered three attackers: one who knows only age and sex, the kind of information found in a medical record or public profile, and two who also know the patient's expression of the 100 or 5,000 most variable genes. The last two are deliberately stringent, because an outside party would rarely hold expression measured on the patient's own sample. A biologically faithful synthetic cohort will necessarily support some population-level inference: if sepsis patients share an expression signature, anyone can use it to guess sepsis status. We therefore compared each attack's success on training patients with its success on held-out patients never seen by the generator; only this excess training success reflects information leaked about individuals [14, 20].

We summarized each attack by the median of the ten largest per-secret excesses, each averaged over folds, so that the most exposed secrets are not diluted by the many gene secrets. This excess was small for the attacker who knew only age and sex (0.028–0.066; at most 0.114–0.134 for any secret) and larger for the 100-gene attacker (0.094–0.143), for whom every secret with an excess above 0.20 was a clinical label (Fig. 5B–D). In each cohort the most exposed secret was the label of the prediction task: tumor stage in TCGA-LUAD (excess 0.518; inferred correctly for 94% of training patients but only 42% of held-out patients), sepsis status (0.192), and disease class in RISK-IBD (0.281). The expression values themselves leaked little (largest per-gene excess 0.092–0.177). Knowing 5,000 genes rather than 100 produced only a moderate further increase (0.167, 0.189, and 0.242), mostly on expression secrets.

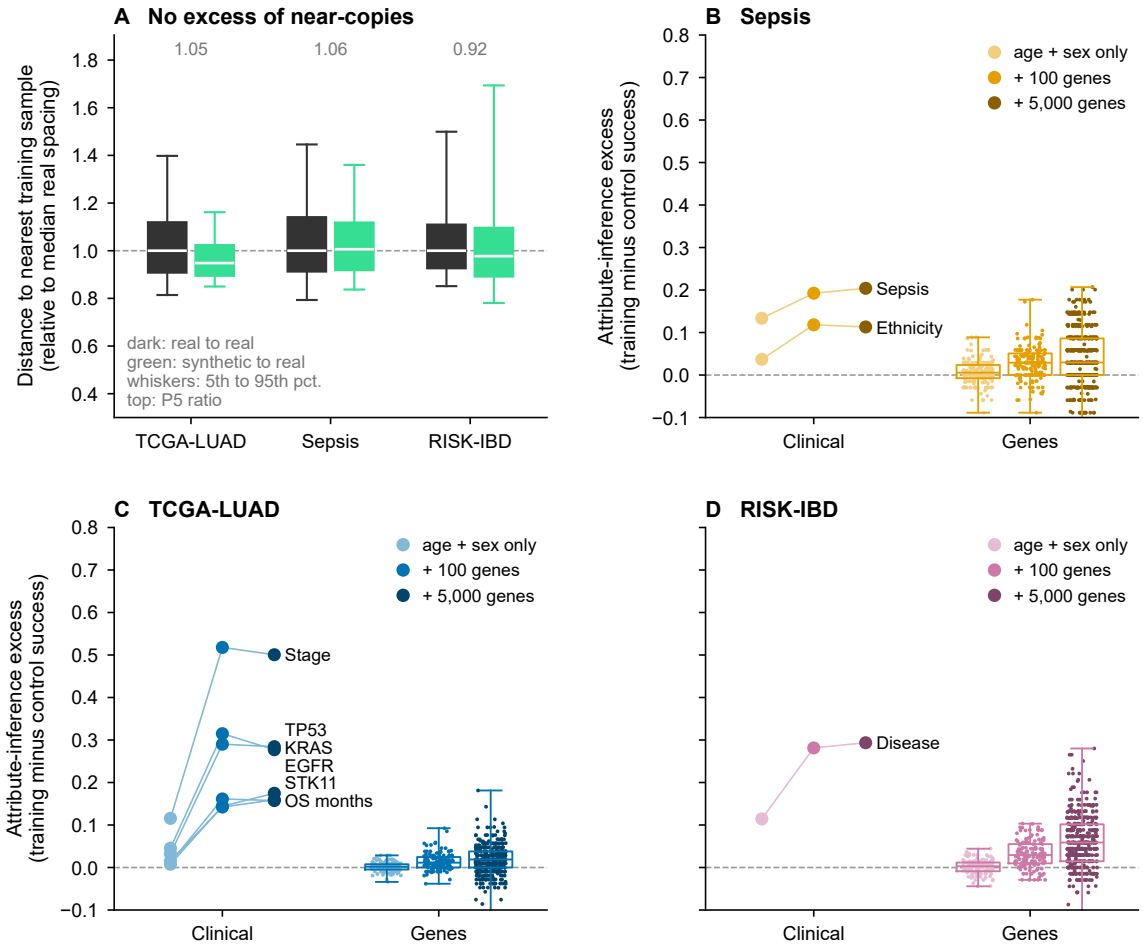


Figure 5. Privacy analyses on the fold-specific synthetic cohorts. Colors in B to D denote cohorts as in Fig. 2. (A) Memorization check. Distance from each synthetic sample to its nearest real training sample (green) and from each real training sample to its nearest other training sample (dark gray), pooled over the five outer folds and scaled by each fold's median real spacing (dashed line). Boxes, quartiles; whiskers, 5th to 95th percentiles. Numbers above give the fifth-percentile ratio (synthetic over real), averaged over folds; values near or above one indicate no excess of near-copies. (B to D) Attribute inference in sepsis (B), TCGA-LUAD (C), and RISK-IBD (D). Excess training success (training minus control attack success) per secret, averaged over folds, for an attacker who knows age and sex (light), age, sex and 100 expression values (mid), or age, sex and 5,000 expression values (dark); dashed line, no excess training success. Clinical secrets are labeled, with lines joining the three attackers; for TCGA-LUAD, only the six with the largest excess averaged over the three attackers are shown (ALK mutation and OS status are omitted). Expression secrets are shown as one dot per gene, with boxes spanning the interquartile range and whiskers the full range; the axis is cut at −0.1, which clips a few 5,000-gene values; for the 5,000-gene attacker they are the genes ranked 5,001 to 5,100 by variance rather than 101 to 200.

Discussion

Across cohorts and analyses, the synthetic cohorts preserved much of the measured structure of the real data and reproduced the main translational analyses we tested: most real DE genes were recovered at close to their real effect sizes; synthetic-trained classifiers performed close to real-trained ones on held-out real patients and relied on largely the same driver genes; and co-expression and clinical structure were retained. Four points follow for how such cohorts should be used: one synthetic cohort can answer many downstream questions, its DE fidelity depends on the statistical power of each contrast, its privacy should be judged against the way it is shared, and it may offer a route around data silos.

One synthetic cohort, many downstream questions

Each released cohort was generated once and then evaluated against several independently defined questions: one to three DE contrasts, a separate prediction task, and per-class SHAP attribution. In TCGA-LUAD, for example, the same synthetic cohort reproduced KRAS- and EGFR-associated differential expression as well as the predictive signal for tumor stage. This matters for translational questions in which several covariates and confounders must each keep their relationship with expression. The classifier results add two practical points. First, because SHAP concordance was measured on real patients excluded from generation, synthetic-trained models can help prioritize genes for follow-up, not only estimate accuracy. Second, the synthetic cohort did not suggest more predictive signal than the real cohort contained: for tumor stage, synthetic-trained models matched the modest real-trained AUC rather than exceeding it.

Which questions a cohort supports depends on the variables declared as anchor columns. The cohorts here used three to five anchor columns, covering a primary phenotype, mutations or subgroups, and age and sex. Questions on undeclared variables are supported less fully. For STK11 mutation, the one variable tested here without being declared, effect sizes shrank but 96.8% of real DE genes kept their direction, suggesting that undeclared variables may still support screening by direction, with weaker significance; more contrasts are needed to establish this. Custodians should therefore declare the variables recipients are most likely to analyze, and each benchmark report lists the anchor columns used.

Differential-expression fidelity depends on real statistical power

In every anchored contrast, synthetic effect sizes matched real ones, yet some real DE genes did not reach synthetic significance. The losses came from genes near the calling thresholds rather than from shrunken effect sizes, so a contrast with many borderline real DE genes, because of small groups, modest effects or high variance, shows lower recovery, as ulcerative colitis did (Supplementary Note 1). The relaxed rule recovers many of these near-misses at some cost in precision, and candidates should be confirmed on real data. DE fidelity is thus not a fixed property of a synthetic dataset but depends on the power of each contrast, which is why each release carries its own benchmark

report.

Interpreting privacy in the context of threat models

We interpret our results in the setting of bilateral sharing of synthetic cohorts: the custodian provides the synthetic cohort to a named recipient under a data-use agreement. Bilateral sharing limits who could mount an attack and adds contractual controls, although it does not change what the synthetic data reveal; the privacy results should therefore be read against this threat model.

Under bilateral sharing, the attribute-inference attacks with 100 and 5,000 known genes are a stringent case. They assume the recipient already holds expression measured on a training patient's own sample, which bilateral sharing does not give them. Even then, for the 100-gene attacker the excess training success was concentrated in clinical labels such as tumor stage, and the expression values themselves leaked little; knowing 5,000 genes rather than 100 produced only a moderate increase, mostly on expression secrets. A recipient who knew only age and sex, the more realistic case, gained little excess training success, and DCR showed no evidence that dbTwin copies training records.

Attribute inference also presupposes that the attacker already knows who the patient is: it measures disclosure of an attribute, not re-identification. The attributes at risk here are clinical variables, such as tumor stage, disease class and mutation status. These variables are not among the identifiers removed under the HIPAA Safe Harbor standard and are routinely retained in de-identified releases, including the public source data for the cohorts studied here. Their inference can still matter when the attacker knows the patient's identity, so custodians should weigh it against the sensitivity of each variable.

Synthetic cohorts as a route around data silos

Much of the clinico-transcriptomic data with the greatest translational value sits in silos: hospital and academic cohorts, biopharma, biobank and consortium datasets, each governed by its own access rules. Moving data between them usually requires data-use agreements, institutional review, and legal negotiation, which slows exploratory work whose payoff is uncertain [1, 2]. Synthetic cohorts released with their benchmark reports could offer a lighter first step. The data custodian generates a synthetic cohort, inspects its benchmark report, and shares it; recipients run exploratory DE, biomarker, and feature-prioritization analyses on it and bring candidate findings to a formal access request for confirmation on the real cohort. The same pattern could serve several settings: a biopharma team could screen a partner's synthetic cohort for target expression, subgroup structure, or biomarker signal before committing to a formal data agreement; a custodian could share one synthetic version with many collaborators and reserve real-data access for the hypotheses that hold up; and methods developers could build and test pipelines on realistic data without waiting for access. In each case, the synthetic cohort accompanied by its benchmark report could speed up the exploratory stage while confirmatory analysis stays with the governed real data.

Limitations

Several limitations should be noted. First, we evaluated only three cohorts, all bulk RNA-seq from single-platform, relatively homogeneous processing pipelines; extension to pseudo-bulk and single-cell data will be important next steps. Second, our predictive models used a single interpretable classifier family (elastic-net logistic regression); whether similar concordance holds for tree-based or deep-learning classifiers remains an open question. Third, DE was limited to six predefined contrasts, and we did not evaluate pathway enrichment or survival models. Fourth, the privacy analyses evaluate scoped threat models. Attribute inference used three attackers, one knowing only age and sex and two also knowing 100 or 5,000 auxiliary genes, each targeting 100 secret genes; it is not exhaustive and does not cover linkage to external genotype resources. Future work should benchmark the privacy of dbTwin cohorts more realistically and comprehensively. Fifth, the benchmark tested only dbTwin; future work will evaluate other synthetic-data generators on the same benchmark for a unified comparison. Finally, the synthetic cohorts are intended for exploratory analysis, hypothesis generation, method development, and preliminary feature prioritization, followed by confirmation on appropriately governed real data. They should not be treated as unquestioned replacements for source data in clinical, regulatory, or other high-stakes applications.

The datasets, fold-level benchmark cohorts, benchmark code, and finished reports provide a reusable test bed for synthetic transcriptomic research. Because the benchmark is generator-agnostic, other groups can apply it to their own generators and compare their results directly with these releases. More broadly, they illustrate a release pattern in which a synthetic dataset is accompanied by the exact evidence needed to judge what it preserves, and what it does not preserve.

Methods

Software availability

Benchmark analyses were implemented in Python using NumPy, SciPy, scikit-learn, PyDESeq2, SHAP, UMAP, and Anonymeter. The benchmark code and completed reports are openly available at https://github.com/Nanda-Aditya/rna-twin, together with the public real datasets. The synthetic datasets are freely and publicly available at https://huggingface.co/datasets/NandaAditya/rna-twin; access is granted automatically on accepting a click-through data-use agreement.

Reproducible cohort-specific benchmark reports

For each source dataset, the release includes one full synthetic cohort, generated from the complete real cohort, the five fold-specific synthetic cohorts used for the held-out analyses, and a benchmark report. Each report details the composition of the real and synthetic cohorts, global expression structure, every DE contrast with per-gene results, fold-level TSTR and SHAP summaries, DCR and attribute-inference results, a data dictionary, and limitations.

Data availability

The full synthetic cohorts and the fold-specific benchmark cohorts for TCGA-LUAD, GSE189400, and GSE57945 are freely and publicly available at https://huggingface.co/datasets/NandaAditya/rna-twin; access is granted automatically on accepting a click-through data-use agreement. All real data, split manifests, and completed benchmark reports are available in the GitHub repository. The source cohorts are publicly available from TCGA (TCGA-LUAD) and from the Gene Expression Omnibus under accessions GSE189400 (whole-blood subseries of SuperSeries GSE189403) and GSE57945. Every benchmark statistic reported here can be recomputed from the released real cohorts, synthetic cohorts, split manifests and benchmark code.

Source cohorts and preprocessing

We used three public studies. TCGA-LUAD provided lung adenocarcinoma RNA-seq expression as RSEM gene-level estimates, together with harmonized clinical annotations and somatic mutation calls, all from the cBioPortal [21, 22] TCGA PanCancer Atlas study [15, 16]. The analysis cohort contained 510 samples, 19,719 expression features, mutation indicators, age, sex, overall-survival fields, and tumor stage; genes with a total of 10 or less across samples were removed. GSE189400 [17], the whole-blood (PAXgene) RNA-seq subseries of SuperSeries GSE189403 from the EARLI cohort of critically ill adults, provided integer read counts for 27,097 genes in 221 subjects, with age, sex, ethnicity, and sepsis classification. We kept the 152 subjects classified as sepsis with bloodstream infection (60) or no sepsis (92), excluded 69 with sepsis from a non-bloodstream source, and removed no genes. GSE57945 [18], from the pediatric RISK inception cohort [19], provided 322 ileal biopsy samples, with 43,790 genes obtained as raw counts from Expression Atlas (E-GEOD-57945) [23], together with disease classification (Crohn's disease, ulcerative colitis, or non-IBD control), age at diagnosis, sex, and sampled organ part. Clinical metadata and expression matrices were joined by sample identifier. Sepsis and RISK-IBD expression are integer read counts. TCGA-LUAD RSEM values are not counts and are non-integer; they were used as provided and, as for every cohort by default, rounded to integers only for DESeq2. All released synthetic expression values are non-negative integers.

dbTwin synthetic generation

dbTwin (dbTwin, Inc.) is a proprietary ML-based generator adapted to clinico-transcriptomic data [11] that operates directly on clinical and gene-expression tables at full transcriptome scale. dbTwin takes two required inputs, expression and metadata tables in CSV or TSV format, and an optional list of up to eight anchor columns : translationally meaningful clinical variables, chosen by the user, on which fidelity is prioritized. For the current release, the anchor columns were EGFR mutation status, KRAS mutation status, age, sex, and tumor stage for TCGA-LUAD; sepsis status, age, and sex for GSE189400; and disease class, age at diagnosis, and sex for RISK-IBD.

Generation and cross-validation design

We separated benchmark generation from release generation. A deterministic five-fold outer manifest assigned each subject to one held-out fold, stratified by the prediction task. For each split, dbTwin was fit only on the remaining real training subjects and generated the same number of synthetic records as the training fold. Held-out expression and clinical values were never used to fit the corresponding generator. In TCGA-LUAD, 2 samples with missing tumor stage, the prediction target, were kept in the full cohort but excluded from the fold design. TCGA-LUAD folds contained 406–407 training and 101–102 held-out subjects; sepsis folds contained 121–122 training and 30–31 held-out subjects; RISK-IBD folds contained 257–258 training and 64–65 held-out subjects.

TSTR, SHAP, DCR, and attribute-inference estimates use these fold-specific cohorts. For the public full-cohort release, the generator was rerun on the complete source cohort and produced a synthetic cohort with the same sample count. Clinical fidelity, co-expression, UMAP, and differential expression were evaluated once on this full synthetic cohort against the complete real cohort. The full release is intended for use; together with the fold-specific cohorts, it allows every benchmark statistic to be independently reconstructed.

Global statistical fidelity

Clinical fidelity was assessed within each prediction class by comparing anchor-column distributions between the real and synthetic cohorts with total variation distance. The anchor column that defines the prediction class (tumor stage, sepsis status, or disease class) is constant within a class and was not compared, leaving 12, 4, and 6 class-by-anchor comparisons in TCGA-LUAD, sepsis, and RISK-IBD. Each column was treated as categorical or continuous once, from its values in the full real cohort. Categorical variables kept missing values as their own level; continuous variables were binned into 10 equal-width intervals defined on the real class, with the outer bins left open. As a sampling-noise reference, we drew 500 bootstrap resamples of each real class with the size of the corresponding synthetic class and computed the same TVD between the real class and each resample; observed values above the 97.5th percentile were counted as exceeding sampling noise.

All expression analyses except DE used $\log_2$ counts per million plus one, $\log_2(\mathrm{CPM} + 1)$, with CPM computed from each sample's total counts over all genes; variable genes were ranked by their variance across samples on this scale. Per-gene statistics in the full real and synthetic cohorts were the mean and the variance across samples. For the sex-chromosome check, we used a fixed panel of XIST and nine Y-linked genes (RPS4Y1, DDX3Y, KDM5D, UTY, EIF1AY, USP9Y, ZFY, NLGN4Y, and TXLNGY), kept the genes present in each cohort (9, 9, and 10 in TCGA-LUAD, sepsis, and RISK-IBD), computed the male-minus-female difference in mean $\log_2(\mathrm{CPM} + 1)$ of each gene in the real and synthetic cohorts, and fitted the least-squares slope of synthetic on real differences. For co-expression, the 500 genes with highest real-cohort variance were kept. Pearson gene–gene correlation matrices were computed separately in real and synthetic data, ordered by average-linkage hierarchical clustering of the real matrix, and compared by Spearman correlation of their upper-triangle entries. For UMAP, genes at or below the 40th percentile of real-cohort

variance were excluded, real and synthetic samples were reduced jointly to 100 principal components (centered, not scaled), and these were embedded in a single two-dimensional UMAP fit [24] (15 neighbors, minimum distance 0.1, Euclidean distance, random state 42). Maximum mean discrepancy (MMD) [25] with an RBF kernel was computed on the two UMAP coordinates.

Differential-expression validation

Differential expression was evaluated once on the full real and full synthetic cohorts. DESeq2 models were fit independently to expression values rounded to integers in the real and synthetic cohorts using the prespecified contrast variable as the design factor [26, 27]; as in R DESeq2, counts flagged as outliers by Cook's distance were replaced and those genes refit, and fold changes were not shrunk. All genes were tested in TCGA-LUAD (19,719), sepsis (27,097) and RISK-IBD (43,790). DE summaries were restricted to the 80% of tested genes with highest mean CPM in the full real cohort (15,775, 21,677, and 35,032 genes). As in the benchmark reports, mutation contrasts are reported as wild-type (0) versus mutant (1), so a positive $\log_2$ fold change means higher expression in wild-type tumors. Real DE genes were defined by adjusted $p < 0.05$ and $|\log_2 \mathrm{FC}| > 0.5$. Synthetic DE genes were called with the same strict rule or with a relaxed rule (adjusted $p < 0.075$, $|\log_2 \mathrm{FC}| > 0.4$). For each rule we recorded the number of synthetic DE genes, their overlap with the real DE set, recall (overlap divided by the real DE count), and precision (overlap divided by the synthetic DE count). Pearson correlation and the least-squares slope of synthetic on real $\log_2$ fold change were computed over all real DE genes, regardless of synthetic significance, so that they test whether real effect sizes are reproduced; a slope below one indicates shrinkage toward zero.

Machine-learning validation

One prediction task was prespecified per cohort: three-class tumor stage in TCGA-LUAD, sepsis status in GSE189400, and disease class (Crohn's disease, ulcerative colitis, or control) in RISK-IBD. Counts were transformed to $\log_2(\mathrm{CPM} + 1)$. Within each outer training fold, genes were ranked by training-set variance and the number retained was 90% of the training sample count, capped at the available gene count. Missing numeric values were median-imputed and features standardized using training-fold statistics.

We used elastic-net logistic regression with balanced class weights, the SAGA solver (random state 42), and fixed ℓ_1 ratio 0.05. Regularization C was evaluated at 51 log-spaced values from 10^{-4} to 10^{1} using stratified inner cross-validation restricted by the smallest class. A candidate was eligible only when the active-feature fraction remained between 20% and 60% in every inner fold and outer split. The eligible value with the highest fold-weighted mean AUC was selected, with smaller C preferred on ties, and then applied unchanged to real- and synthetic-trained models. The selected values were $C = 0.005$, 0.016, and 0.0063 for TCGA-LUAD, sepsis, and RISK-IBD.

TRTR and TSTR classifiers were evaluated on the same held-out real subjects. Binary ROC-AUC was used for sepsis; TCGA-LUAD and RISK-IBD used mean one-versus-rest AUC. Linear

SHAP values [28] were computed on standardized held-out real features. For each class and split, mean absolute SHAP values from real- and synthetic-trained models were compared over the 50 genes with greatest real-trained importance using Pearson and Spearman correlation. Correlations were defined only when both importance vectors were finite and had positive standard deviation; undefined folds were excluded from the mean and standard deviation. Top-10 agreement was the fraction of the ten most important real-trained genes that were also among the ten most important synthetic-trained genes.

To test sensitivity to regularization, we refit real- and synthetic-trained models in every outer fold at 28 values of C from the same grid (0.0005 to 0.25), with the ℓ_1 ratio, gene selection, and preprocessing unchanged. At each C we recorded TRTR and TSTR AUC, the fraction of input genes with any non-zero real-trained coefficient, and SHAP Pearson r over the top 50 real-trained genes. Because inactive genes have zero attribution in both models and can inflate r, we also computed r over the top $\min(50, n_{\text{active}})$ genes. Refitting at the selected C reproduced the benchmark AUC to within 0.01 and SHAP r to within 0.02 in every fold.

Distance-based memorization analysis

DCR was treated as a distance-based memorization metric, not as privacy proof. All genes were transformed to $\log_2(\text{CPM} + 1)$ and standardized using the real training mean and standard deviation. For each split, we computed the Euclidean distance from every synthetic record to its nearest real training record, $d_{S\to T}$. As the reference distribution, we computed each real training record's distance to its second-nearest training neighbor, $d^{(2)}_{T\to T}$; the second neighbor excludes the zero self-distance. The scalar statistic was

$$\text{DCR}_{p5} = \frac{P_5(d_{S\to T})}{P_5(d^{(2)}_{T\to T})}, \tag{1}$$

where P_5 denotes the fifth percentile; the median ratio was computed analogously. Values near one indicate that the closest synthetic-to-training distances resemble ordinary nearest-neighbor spacing within the training cohort; values well below one indicate that synthetic samples sit closer to training records than training records sit to each other, as expected from memorization.

Attribute-inference evaluation

Attribute inference was implemented with Anonymeter's inference evaluator at 95% confidence and one attack per held-out subject [20]. The original real training fold was the attack population, its synthetic counterpart was the released information, and the held-out real fold was the matched control. Clinical quasi-identifiers were age and sex. Eligible clinical secrets were all remaining configured clinical fields. Expression auxiliaries were selected by synthetic-cohort variance and used as $\log_2(\text{CPM} + 1)$ standardized with training statistics; expression secrets were the next non-overlapping variance-ranked genes, attacked on their released raw counts. All cohorts used 100 secret genes. Each secret was attacked by three attackers: one with the clinical quasi-identifiers

alone, and two that also held the top 100 or the top 5,000 variance-ranked genes as expression auxiliaries. For the 5,000-gene attacker, the expression secrets were the next 100 genes (ranks 5,001–5,100), with the same evaluator, folds, quasi-identifiers, and seeds.

Anonymeter performed classification for categorical or binary secrets and regression for continuous numeric secrets. For each secret and split we retained the training attack success rate r_{train} and matched-control success rate r_{control}. For each secret, $r_{\text{train}} - r_{\text{control}}$ was averaged over splits; we refer to this difference as the excess training success. As in the benchmark reports, the primary summary for each attack was the median of the ten largest of these per-secret means. We selected this unnormalized difference because it directly reports the attacker's advantage on training patients [14]. The normalized Anonymeter estimator divides by $1 - r_{\text{control}}$ and can therefore assign a larger value when the same absolute excess success occurs against a higher-performing control. Both quantities remain available in the machine-readable artifacts. For Fig. 5B–D, excess was averaged over folds per secret and secrets were grouped as clinical or expression secrets, with only the six TCGA-LUAD clinical secrets of largest excess averaged over the three attackers shown; secrets that are constant in the real cohort were omitted.

Use of AI tools

AI tools were used for language editing and to assist in writing analysis code. The author reviewed all code, outputs, interpretations and references, and takes full responsibility for the content.

Ethics approval and consent to participate

This study is a secondary analysis of publicly available, de-identified data from TCGA-LUAD, GEO GSE189400 and GEO GSE57945. No new data were collected from human participants, and the author had no access to identifying information. Ethics approval and informed consent for the source cohorts were obtained by the original studies, as described in their publications [15, 17–19]. Because only publicly available, de-identified data were analyzed, this work did not require separate institutional review board approval.The released synthetic cohorts contain no real patient records.

Consent for publication

Not applicable.

Author Contributions

AN: Conceptualization, methodology, software implementation, data acquisition and preprocessing, development and implementation of the generative model, benchmarking and evaluation protocols, literature review, and principal manuscript drafting.

Funding

This work was carried out at dbTwin, Inc. and received no external funding.

Acknowledgments

AN thanks Ashish Patel for his support and help. We thank the cBioPortal, TCGA, Gene Expression Omnibus, and Expression Atlas teams for providing accessible RNA-seq data resources.

Competing Interests

AN is a co-founder of and owns equity in dbTwin, Inc., which is developing the tool described in this manuscript.

References

[1] James Brian Byrd, Anna C. Greene, Deepashree Venkatesh Prasad, Xiaoqian Jiang, and Casey S. Greene. Responsible, practical genomic data sharing that accelerates research. *Nature Reviews Genetics*, 21(10):615–629, 2020. doi: 10.1038/s41576-020-0257-5.

[2] Zhiyu Wan, James W. Hazel, Ellen Wright Clayton, Yevgeniy Vorobeychik, Murat Kantarcioglu, and Bradley A. Malin. Sociotechnical safeguards for genomic data privacy. *Nature Reviews Genetics*, 23(7):429–445, 2022. doi: 10.1038/s41576-022-00455-y.

[3] Eric E. Schadt, Sangsoon Woo, and Ke Hao. Bayesian method to predict individual SNP genotypes from gene expression data. *Nature Genetics*, 44(5):603–608, 2012. doi: 10.1038/ng.2248.

[4] Arif Harmanci and Mark Gerstein. Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nature Methods*, 13(3):251–256, 2016. doi: 10.1038/nmeth.3746.

[5] Gamze Gürsoy, Tianxiao Li, Susanna Liu, Eric Ni, Charlotte M. Brannon, and Mark B. Gerstein. Functional genomics data: privacy risk assessment and technological mitigation. *Nature Reviews Genetics*, 23(4):245–258, 2022. doi: 10.1038/s41576-021-00428-7.

[6] Shuvom Sadhuka, Daniel Fridman, Bonnie Berger, and Hyunghoon Cho. Assessing transcriptomic reidentification risks using discriminative sequence models. *Genome Research*, 33(7):1101–1112, 2023. doi: 10.1101/gr.277699.123.

[7] Sanaz Kavianpour, James Sutherland, Esma Mansouri-Benssassi, Natalie Coull, and Emily Jefferson. Next-generation capabilities in trusted research environments: interview study. *Journal of Medical Internet Research*, 24(9):e33720, 2022. doi: 10.2196/33720.

[8] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, Sébastien Ourselin, Micah Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and M. Jorge Cardoso. The future of digital health with federated learning. *npj Digital Medicine*, 3:119, 2020. doi: 10.1038/s41746-020-00323-1.

[9] The-Chuong Trinh, Jean-Baptiste Woillard, Guido Uguzzoni, and Christophe Battail. A unified benchmark of synthetic data generation for clinical transcriptomic cancer cohorts. *bioRxiv*, 2026. doi: 10.64898/2026.05.13.724858.

[10] Hakime Öztürk, Tejumade Afonja, Joonas Jälkö, Ruta Binkyte, Pablo Rodriguez-Mier, Sebastian Lobentanzer, Andrew Wicks, Jules Kreuer, Sofiane Ouaari, Nico Pfeifer, Shane Menzies, Sikha Pentyala, Daniil Filienko, Steven Golob, Patrick McKeever, Jineta Banerjee, Luca Foschini, Martine De Cock, Julio Saez-Rodriguez, Mario Fritz, Oliver Stegle, and Antti Honkela. Towards useful and private synthetic omics: Community benchmarking of generative models for transcriptomics data. *bioRxiv*, 2026. doi: 10.64898/2026.03.02.707794.

[11] Aditya Nanda and Somdutta Saha. Synthetic RNA-seq cohorts for data sharing: a discovery-aware benchmark at transcriptome scale. *bioRxiv*, May 2026. doi: 10.64898/2026.05.22.726357.

[12] Chao Yan, Yao Yan, Zhiyu Wan, Ziqi Zhang, Larsson Omberg, Justin Guinney, Sean D. Mooney, and Bradley A. Malin. A multifaceted benchmarking of synthetic electronic health record generation models. *Nature Communications*, 13(1):7609, 2022. doi: 10.1038/s41467-022-35295-1.

[13] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data–anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1451–1468, 2022.

[14] Lisa Pilgram, Fida K. Dankar, Jörg Drechsler, Mark Elliot, Josep Domingo-Ferrer, Paul Francis, Murat Kantarcioglu, Linglong Kong, Bradley Malin, Krishnamurty Muralidhar, Puja Myles, Fabian Prasser, Jean Louis Raisaro, Chao Yan, and Khaled El Emam. A consensus privacy metrics framework for synthetic data. *Patterns*, 6(10):101320, 2025. doi: 10.1016/j.patter.2025.101320.

[15] Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–550, 2014. doi: 10.1038/nature13385.

[16] Jianfang Liu, Tara Lichtenberg, Katherine A. Hoadley, Laila M. Poisson, Alexander J. Lazar, Andrew D. Cherniack, Albert J. Kovatich, Christopher C. Benz, Douglas A. Levine, Adrian V. Lee, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416.e11, 2018. doi: 10.1016/j.cell.2018.02.052.

[17] Katrina L. Kalantar, Lucile Neyton, Mazin Abdelghany, Eran Mick, Alejandra Jauregui, Saharai Caldera, Paula Hayakawa Serpa, Rajani Ghale, Jack Albright, Aartik Sarma, et al. Integrated host-microbe plasma metagenomics for sepsis diagnosis in a prospective cohort of critically ill adults. *Nature Microbiology*, 7(11):1805–1816, 2022. doi: 10.1038/s41564-022-01237-2.

[18] Yael Haberman, Timothy L. Tickle, Phillip J. Dexheimer, Mi-Ok Kim, Dora Tang, Rebekah Karns, Robert N. Baldassano, Joshua D. Noe, Joel Rosh, James Markowitz, et al. Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. *Journal of Clinical Investigation*, 124(8):3617–3633, 2014. doi: 10.1172/JCI75436.

[19] Subra Kugathasan, Lee A. Denson, Thomas D. Walters, Mi-Ok Kim, Urko M. Marigorta, Melanie Schirmer, Kajari Mondal, Chunyan Liu, Anne Griffiths, Joshua D. Noe, et al. Prediction of complicated disease course for children newly diagnosed with Crohn's disease: a multicentre inception cohort study. *The Lancet*, 389(10080):1710–1718, 2017. doi: 10.1016/S0140-6736(17)30317-3.

[20] Matteo Giomi, Franziska Boenisch, Christoph Wehmeyer, and Borbála Tasnádi. A unified framework for quantifying privacy risk in synthetic data. *Proceedings on Privacy Enhancing Technologies*, 2023(2):312–328, 2023. doi: 10.56553/popets-2023-0055.

[21] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E. Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J. Byrne, Michael L. Heuer, Erik Larsson, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5):401–404, 2012. doi: 10.1158/2159-8290.CD-12-0095.

[22] Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S. Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, 6(269):pl1, 2013. doi: 10.1126/scisignal.2004088.

[23] Pablo Moreno, Silvie Fexova, Nancy George, Jonathan R. Manning, Zhichiao Miao, Suhaib Mohammed, Alfonso Muñoz-Pomer, Anja Fullgrabe, Yalan Bi, Natassja Bush, et al. Expression Atlas update: gene and protein expression in multiple species. *Nucleic Acids Research*, 50(D1): D129–D140, 2022. doi: 10.1093/nar/gkab1030.

[24] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. doi: 10.48550/arXiv.1802.03426.

[25] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

[26] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014. doi: 10.1186/s13059-014-0550-8.

[27] Boris Muzellec, Maria Teleńczuk, Vincent Cabeli, and Mathieu Andreux. PyDESeq2: a Python package for bulk RNA-seq differential expression analysis. *Bioinformatics*, 39(9):btad547, 2023. doi: 10.1093/bioinformatics/btad547.

[28] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- si.pdf