


Optimizing sample size for supervised machine learning with bulk transcriptomic sequencing: a learning curve approach

Yunhui Qi^{1,2}, Xinyi Wang^{1,3}, Li-Xuan Qin ^{1,*}

¹Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 633 Third Avenue, New York, NY 10017, United States

²Department of Statistics, Iowa State University, 2438 Osborn Drive, Ames, IA, 50011-1090, United States

³Department of Statistics, The University of California, 1 Shields Ave, Davis, CA 95616, United States

*Corresponding author. Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 633 Third Avenue, New York, NY 10017, United States. E-mail: qinl@mskcc.org

Abstract

Accurate sample classification using transcriptomics data is crucial for advancing personalized medicine. Achieving this goal necessitates determining a suitable sample size that ensures adequate classification accuracy without undue resource allocation. Current sample size calculation methods rely on assumptions and algorithms that may not align with supervised machine learning techniques for sample classification. Addressing this critical methodological gap, we present a novel computational approach that establishes the accuracy-versus-sample size relationship by employing a data augmentation strategy followed by fitting a learning curve. We comprehensively evaluated its performance for microRNA and RNA sequencing data, considering diverse data characteristics and algorithm configurations, based on a spectrum of evaluation metrics. To foster accessibility and reproducibility, the Python and R code for implementing our approach is available on GitHub. Its deployment will significantly facilitate the adoption of machine learning in transcriptomics studies and accelerate their translation into clinically useful classifiers for personalized treatment.

Keywords: sample size; machine learning; transcriptomics; bulk sequencing

Introduction

Accurate sample classification using transcriptomic sequencing data is pivotal for guiding personalized treatment decisions [1–6]. The success of such endeavors depends on the selection of an appropriate sample size, to achieve adequate statistical power while avoiding undue resource allocation or ethical concerns [7–12]. Various sample size calculation methods are available to identify differentially expressed markers [13–19]. These methods establish connections between the required sample size, the desired power, and the projected effect size within a hypothesis-testing framework, employing either closed-form formulae derived from statistical tests [13–16] or *in silico* simulations based on parametric distributions [17–19]. When the study goal shifts to developing a multimarker classifier, fewer sample size calculation methods are available. They were primarily developed for microarray data and, in principle, can be adapted for sequencing data [20–24]. These methods establish relationships between the required sample size and the desired classification accuracy, through either formulae derived from parametric distributions [20–22] or simulations via subsampling [23, 24]. However, none of these methods are compatible with modern supervised machine learning techniques, as these techniques eschew parametric distribution assumptions and require a substantial number of

samples, making subsampling infeasible [25–28]. Consequently, there is a pressing need to develop sample size calculation methods compatible with machine learning in classification studies using transcriptomic sequencing data.

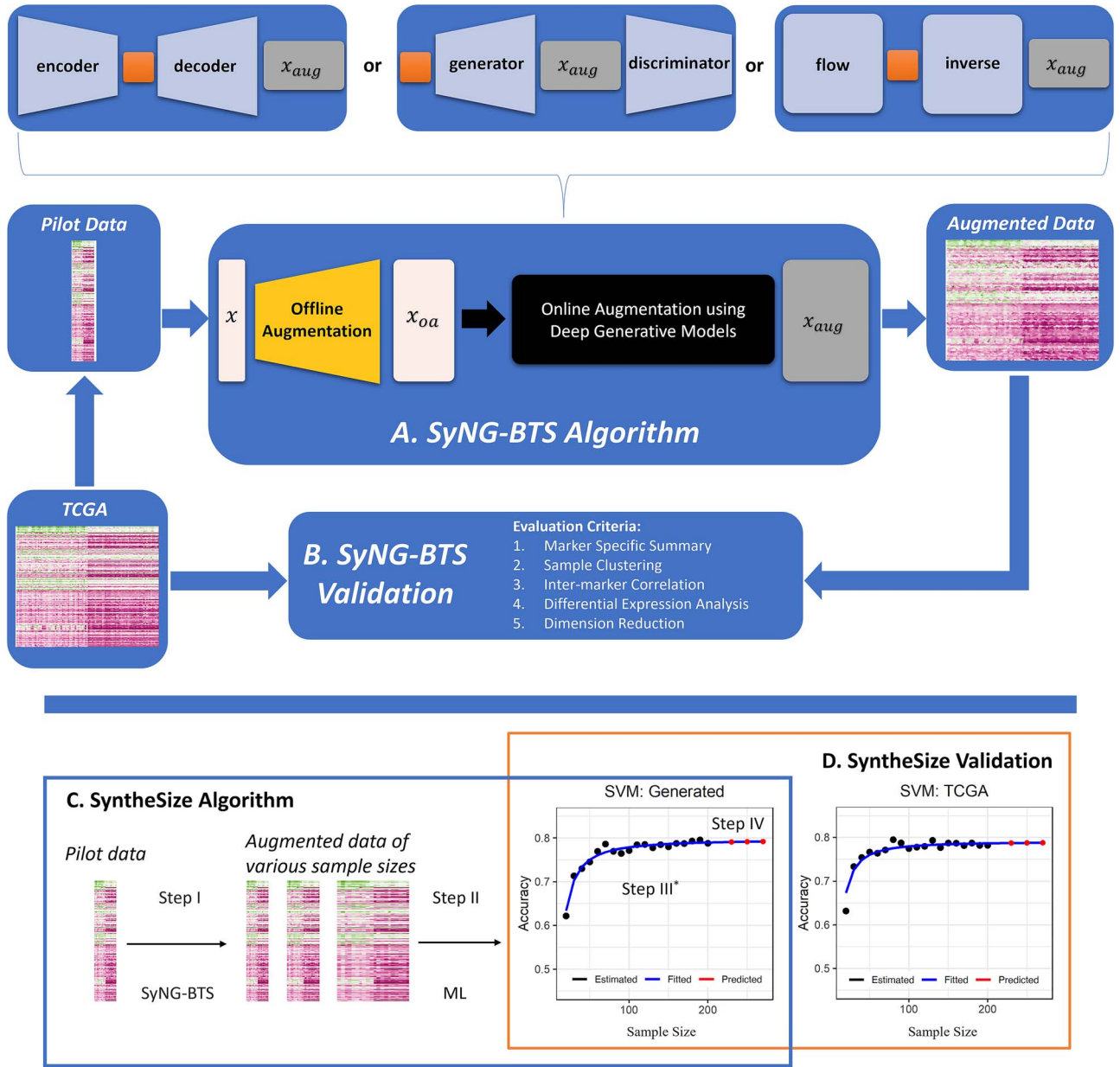
We developed a new computational approach to fill this methodology gap. Our approach entails two stages: first, synthesizing realistically distributed transcriptomic sequencing data without relying on a predefined formula, and second, determining a suitable sample size based on the synthesized data across a range of sample sizes. Specifically, we (i) build data augmentation tools that harness the power of deep generative models (DGMs), which will be trained on available pilot data and subsequently used to generate data for any desired number of samples [29–31] and (ii) ascertain a suitable sample size by fitting the inverse power law function (IPLF) with augmented data across different sample sizes and their respective classification accuracies using a machine learning technique [32, 33] (Fig. 1). We name our algorithm for the first stage SyNG-BTS (pronounced “sing-beats”), representing Synthesis of Next-Generation Bulk Transcriptomic Sequencing, and the algorithm for the second stage SynthesizeSize.

DGMs are designed to simulate data resembling real-world observations, which can be especially useful when acquiring real data is challenging [34–37]. DGMs initially received acclaim for augmenting imaging data [34] and recently achieved successes in

Received: September 11, 2024. Revised: January 11, 2025. Accepted: February 21, 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com



* Step III (IPLF): $Accuracy = f(n; a, b, c) = (1 - a) - bn^c$; $a \in [0, 1]$, $b \in [0, \infty)$, $c \in [-1, 0]$;

The values of parameters a , b , and c depend on the characteristics of the dataset and the learning capacity of the chosen learning method.

Figure 1. Schema of SyNG-BTS and Synthesize algorithms along with their respective validation. (A) SyNG-BTS algorithm. A pilot dataset, denoted as x , undergoes offline augmentation to transform into x_{0a} , which subsequently serves as the input for a DGM, leading to augmentation to x_{aug} . (B) SyNG-BTS validation. Pilot datasets are generated by randomly sampling from a dataset in TCGA and are subsequently inputted into SyNG-BTS to generate augmented data with the matching sample size of the source TCGA dataset, enabling a comprehensive comparison of the empirical data and the augmented data using various evaluation criteria. (C) Synthesize algorithm. It determines the sample size for machine learning (ML), such as support vector machine (SVM), in four steps: (I) a pilot dataset is augmented with SyNG-BTS and sampled to generate datasets with a range of sample sizes; (II) each augmented dataset is used to train a classifier with a chosen ML technique (such as SVM) and assess its accuracy; (III) the estimated accuracies for various sample sizes are fitted to a learning curve using the IPLF; and (IV) utilizing the fitted curve, the prediction accuracy is projected for any desired sample size. (D) Synthesize validation. ML is used to classify the samples in the TCGA breast cancer study to two subtypes, IDC and ILC. Data are split to two portions: one is used to supply pilot data and derive the IPLF curve with Synthesize, and the other is to provide empirical datasets of various sample sizes via subsampling to construct another IPLF curve for comparison with the Synthesize-derived IPLF curve.

single cell sequencing [35–37]. Several families of DGMs are popularly used for data augmentation, including variational autoencoders (VAEs), generative adversarial networks (GANs), and flow-based generative models [38–41]. When employing these models to simulate bulk tissue transcriptomic sequencing, a challenge arises due to the typically modest sample size of the training data, as its randomness after online augmentation (i.e. augmentation by DGMs in real-time) may have a strong impact on the training

process. To address this issue, we opt for offline augmentation, which generates and stores augmented copies of the original data before training, ensuring a more consistent set of generated samples and fostering more stable learning dynamics. We will utilize an autoencoder (AE) for offline augmentation when dealing with a relatively modest number of markers, such as in microRNA sequencing (miRNA-seq) [42–45]. AEs are designed for reconstructing the input data rather than generating new

samples. In the case of RNA sequencing (RNA-seq), where the number of markers is substantial, we will employ Gaussian white noise addition [46, 47]. It introduces noise generated from a Gaussian distribution to the pilot data and iterates this procedure multiple times. The resulting datasets are aggregated to expand the sample size of the original pilot data.

IPLFs are utilized to represent learning curves that depict the relationship between a classifier's accuracy and the training data's sample size [32, 33]. The term "learning curve" is often used to portray the learning trajectory, illustrating how a learner's performance improves with experience and practice [33]. In the realm of machine learning, a learning curve refers to a graphical representation or a mathematical function that illustrates how a learning model's performance improves with an increasing amount of training data. Such curves establish a connection between the prediction accuracy for a learning technique and the sample size of the training dataset. More specifically, these curves typically adhere to the IPLF, displaying three distinct and sequential phases characterized by (i) rapid performance enhancement, (ii) gradual progress, and (iii) eventual plateauing. The IPLF defines these phases with three parameters: the learning rate, decay rate, and minimum achievable error rate. This inverse-power-law "learning" behavior appears to be widespread and has been observed in diverse prediction contexts [48]. It has been utilized for sample size determination in sample classification with microarray data, initially in an unweighted manner by Mukherjee et al. (2003) [23] and subsequently refined by Figueroa et al. (2012) [24] using a weighted strategy to favor larger sample sizes in model fitting. Here, we employ the method introduced by Figueroa et al. (2012) [24] in conjunction with our data augmentation algorithm to relate a learning technique's accuracy to the sample size of transcriptomic sequencing data that is synthesized via data augmentation.

In this article, we present a comprehensive workflow that leverages the SyNG-BTS algorithm and evaluate its performance in augmenting both miRNA-seq and RNA-seq data. Our investigation delves into critical nuances in algorithm specifications, including model choice, hyperparameter tuning, and offline augmentation, alongside key pilot data characteristics, such as sample size, marker filtering, and data normalization. Performance evaluation is grounded in pilot data sourced from The Cancer Genome Atlas (TCGA), which also serves as reference data for comparison with the augmented data [49–55]. We utilize a spectrum of evaluation metrics, including marker-specific summaries, sample clustering, between-marker correlations, and differential expression analysis [56–60]. Furthermore, we extend this workflow to integrate the SyNG-BTS algorithm and the SyntheSize algorithm and assess its performance in *post-hoc* sample size calculation for TCGA studies. We apply this extended workflow to calculate the sample size for developing predictors of immunotherapy response in a study of advanced clear cell renal carcinoma [61], providing insights into study design using machine learning for immunotherapy outcome prediction. This novel workflow effectively bridges a methodological gap, addressing a critical challenge in the design of transcriptomic sequencing studies using machine learning. Their deployment promises to significantly enhance the likelihood of deriving valuable outcome predictors for personalized treatment of patients.

Methods

Due to space limitations, an overview of the SyNG-BTS and SyntheSize algorithms, along with their performance evaluation and

data application, is provided here. Full details are available in the Supplementary Methods.

Overview of SyNG-BTS

The objective of SyNG-BTS is to train DGMs on a pilot set of bulk transcriptomic sequencing data and subsequently generate data for any number of samples using the trained model (Fig. 1A). Algorithmically, the training of SyNG-BTS involves two main steps, with the first being optional depending on the pilot data characteristics: (i) offline data augmentation using either an AE head or a Gaussian head and (ii) online data augmentation using VAEs, GANs, or flow-based generative models. We also explored their variants, such as conditional VAE (CVAE), Wasserstein GAN (WGAN), WGAN with gradient penalty (WGANGP), masked autoregressive flow (MAF), generative flow with invertible 1×1 convolutions (Glow), and real-valued nonvolume preserving (RealNVP) [29, 30, 62–67]. For all models and their variants, we evaluated values for two shared hyperparameters: the number of learning epochs and the size of learning batches, along with an additional hyperparameter specific to VAE and CVAE [68, 69].

Performance evaluation of SyNG-BTS

We evaluated the performance of SyNG-BTS using data from four TCGA datasets studying skin cutaneous melanoma (SKCM), acute myeloid leukemia (LAML), breast invasive carcinoma (BRCA), and prostate adenocarcinoma (PRAD). These datasets served a dual purpose: (i) acting as sources for subsampling to generate pilot datasets and (ii) serving as reference datasets for assessing the augmented data quality (Fig. 1B). To determine the quality of the augmented data, we assessed its congruence with the reference empirical data using five key metrics. These metrics, detailed in the Supplementary Methods, collectively captured marker-specific, inter-marker, and inter-sample data characteristics in both one-group and two-group settings. For miRNA-seq data augmentation, we examined all four TCGA datasets in the one-group setting, presenting the results for SKCM in the main text and that for the other three in the Supplementary Figures. In the two-group setting, we used the combination of SKCM and LAML datasets (referred to as SKCM/LAML) and the combination of BRCA and PRAD datasets (referred to as BRCA/PRAD), with the results for the latter presented in the Supplementary Figures. For RNA-seq data augmentation, which needs larger pilot data than miRNA-seq due to the considerably greater number of markers, we focused on the BRCA and PRAD RNA-seq datasets for the one-group setting (with the PRAD results presented in the Supplementary Figures) and the BRCA/PRAD combination for the two-group setting.

Overview of SyntheSize

Our proposed approach for sample size determination using augmented datasets is implemented in four main steps (Fig. 1C).

- (I) Data augmentation. Select a set of candidate sample sizes that are evenly distributed (denoted as n_i for $i = 1, \dots, m$) and generate data for each n_i sample size using SyNG-BTS.
- (II) Classifier training. Use each augmented dataset to train a classifier with a chosen learning technique (such as support vector machine) and assess its accuracy. Steps (I) and (II) can be repeated for multiple augmented datasets of each sample size n_i to obtain multiple accuracy estimates, providing a more stable average estimate.
- (III) Learning curve fitting. Fit the estimated accuracies for all candidate sample sizes to a learning curve using the IPLF. Its parameters are estimated via a nonlinear weighted least squares optimization, employing the `nl2sol` routine from the

Port Library, as outlined by Figueroa *et al.* (2012) [24]. In this optimization, the weight for the i -th sample size is i/m , placing greater emphasis on larger sizes.

- (IV) Sample size projection. Utilizing the fitted curve, the prediction accuracy is projected for any desired sample size, which applies the IPLF with the estimated parameters. In particular, the fitted curve can be used to extrapolate the accuracy level for a larger sample size than n_m .

Performance evaluation of SyntheSize

For illustration and evaluation purposes, we applied SyntheSize for *post-hoc* sample size evaluation in the TCGA BRCA study to classify its two subtypes, invasive ductal carcinoma (IDC) versus invasive lobular carcinoma (ILC), using both miRNA-seq and RNA-seq data. More specifically, we initially reserved 100 IDC and 100 ILC samples as an independent validation set, which would be utilized to estimate the *de facto* IPLF of classification accuracy. The remaining samples were then employed to draw pilot data for SyNG-BTS as part of the SyntheSize algorithm. We considered three commonly used classifiers: support vector machine, k-nearest neighbors with $k=20$, and XGBoost. We trained classifiers using generated samples (across a range of candidate sample sizes) and using real samples (over the same candidate sample size range) from the independent validation set. The accuracies of the classifiers at each candidate sample size were utilized to fit the IPLF. Subsequently, the fitted functions derived from generated samples were compared with those from real samples. This analysis provides insights into the effectiveness and reliability of the SyntheSize algorithm for determining sample size in supervised machine learning with transcriptomic sequencing data.

Application of SyntheSize to an immunotherapy study

To further illustrate, we utilized SyntheSize for sample size assessment in predicting patient response to a PD-1 inhibitor, nivolumab, with RNA-seq in advanced clear cell renal cell carcinoma. The objective was to build a classifier with RNA-seq data to distinguish two clinical response groups according to RECIST 1.1: complete or partial response (CR/PR) versus stable or progressive disease (PD/SD), as outlined in the original paper. Pilot data came from a recent study of advanced clear cell renal cell carcinoma involving 152 patients (39 CR/PRs and 113 PD/SDs), all treated with nivolumab and with available RNA-seq data [61].

Results

SyNG-BTS successfully augmented one-group miRNA data

We conducted a comprehensive evaluation of various facets of the augmented data, encompassing marker-specific summary statistics (mean, variation, and sparsity), inter-marker relationships (particularly partial correlation among miRNAs belonging to the same polycistronic clusters), and inter-sample relationships (assessed by clustering the augmented data from SyNG-BTS with the empirical data from TCGA), across the four TCGA studies (Fig. 2; Supplementary Figs S1–S4). In general, the augmented data exhibited high comparability with the empirical data when suitable DGMs and reasonable pilot data sample sizes were utilized, with the latter depending on the specific DGM. The degree of comparability was further influenced by the interplay of pilot data characteristics and algorithm configurations. Detailed results are presented below in the context of the TCGA SKCM study (Fig. 2; Fig. S1), with similar observations noted in the other three TCGA studies (Figs S2–S4).

Model choice played a crucial role in the augmented data quality (Fig. 2). Among the DGMs examined, VAE (specifically with the ratio between reconstruction loss and Kullback–Leibler divergence being 1:10, shorthand as VAE1-10) and flow-models (especially MAF) emerged as the top performers across all evaluation metrics (Fig. 2A–D). Compared with VAEs, MAF better preserved the proportions of expressed markers (i.e. markers with nonzero reads in at least one sample) (Fig. 2C and D). Furthermore, VAE1-10 excelled in scenarios favoring deep training, typically with a fixed number of epochs or a small batch size, while MAF showed relative insensitivity to batch size and performed well with early stopping (Fig. 2E and F). Among the GAN-based models, WGANGP outperformed GAN and WGAN across most of the evaluation metrics especially for marker-specific means and sample clustering (Fig. 2A–D). It also showed overall insensitivity to batch sizes and epoch strategies although occasionally favored early stopping (Fig. 2E and F).

The pilot data characteristic with the most significant impact was sample size (Fig. 2). Increasing the pilot data sample size considerably improved data congruence for VAEs and flow-based models, as evidenced by enhancements across the evaluation metrics (Fig. 2A–D). Take VAE1-10 as an example, as the pilot data sample size increased from 20 to 100, the similarity of marker-specific means and standard deviations greatly improved, nearly halving the median absolute deviation between the augmented data and empirical data (Fig. 2A and B); the mixing of the two data sources upon clustering sharply enhanced, raising the complimentary Adjusted Rand Index (cARI) from about 0.55 to nearly 1; the concordance of inter-marker correlations gradually increased, with the correlation coefficient rising from 0.73 to 0.80 (Fig. 2C). On the other hand, the GAN family performed poorly across all pilot data sample sizes, especially in terms of marker-specific summary statistics (Fig. 2A and B). Hence, a reasonable sample size (40 or more for VAE1-10 and 60 or more for MAF) proved effective for model training, a phenomenon particularly pronounced for high-performing models like VAEs and flow-based models.

In addition to the pilot data sample size, we evaluated the impact of marker filtering (Fig. 2B versus Fig. 2A, Fig. 2D versus Fig. 2C, and Fig. 2E–H versus Fig. S1) and sequencing depth normalization (Fig. 2G and Fig. S1C) for pilot data on the efficacy of model training. The effectiveness of marker filtering (i.e. removing markers with consistently low expression across samples) was evident, leading to a substantial enhancements in both the nonzero marker proportions (with its difference between the augmented data and empirical data decreasing from ~25% to ~5% for VAE1-10 and from ~15% to 0% for MAF) and the mixing of samples from the two data sources (with the cARI increasing from about 0.55 to 0.95 for VAE1-10 using 40 pilot samples and from around 0.92 to 0.99 for MAF using 60 pilot samples). Its impact on the inter-marker correlation metric varied depending on the model, with notable improvements for flow-based models. The use of depth normalization may or may not improve data congruence in the one-group setting. Although trimmed mean of M-values (TMM) and total-count normalization outperformed upper-quartile normalization, they were found to be roughly equivalent or slightly inferior to no normalization.

We further evaluated the impact of offline augmentation on model training (Fig. 2H). Offline augmentation via AE reconstruction proved effective in facilitating the training process, resulting in further enhancement even for the top-performing models like VAE1-10 and MAF. However, it did not improve the performance of GAN models, underscoring the challenges of training GANs in this context.

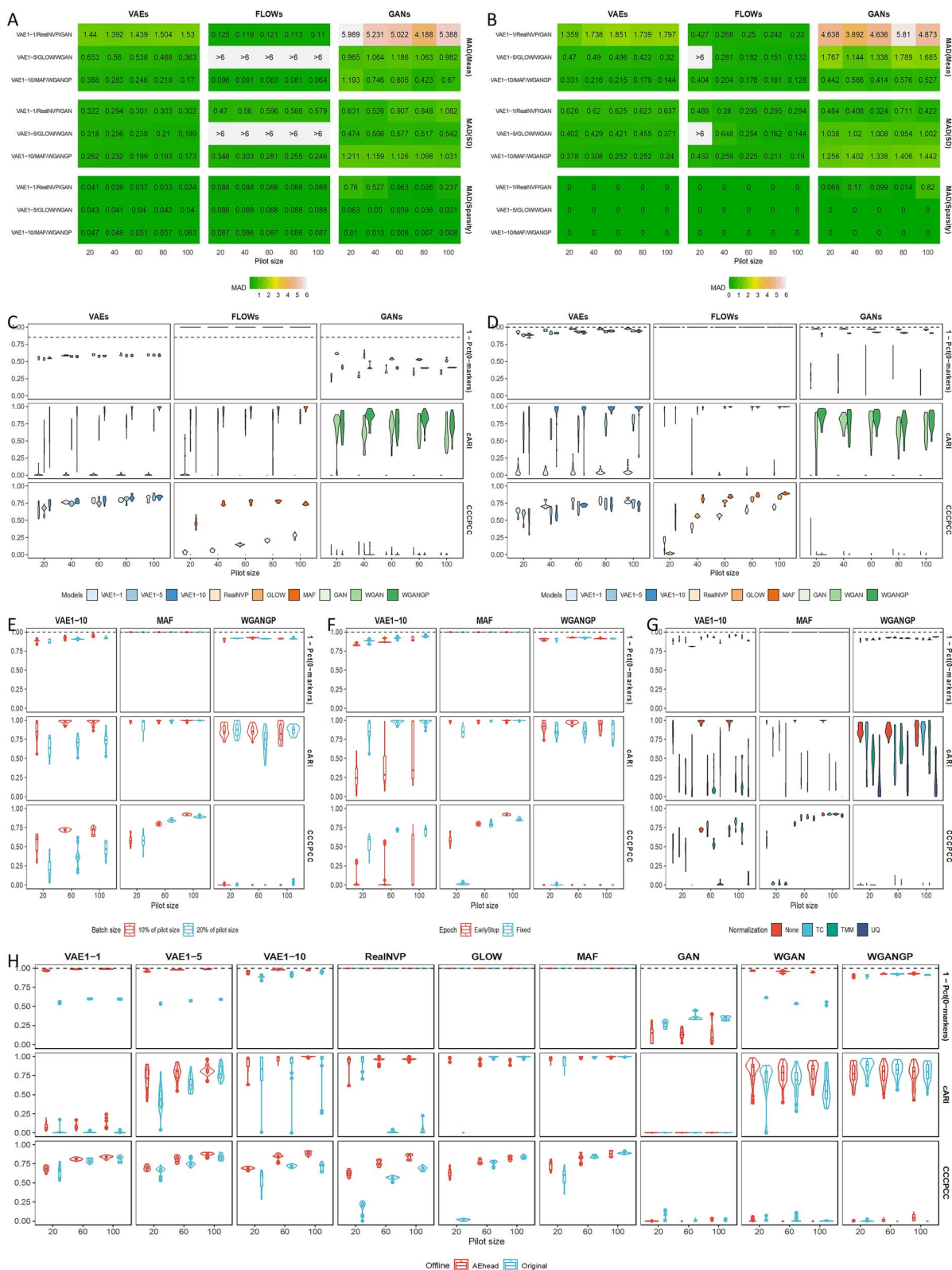


Figure 2. SyNG-BTS evaluation for microRNA-seq in the one-group setting, using pilot data from the TCGA SKCM study without marker filtering (Panels A and C) and with marker filtering (Panels B and D–H). (A) Median absolute deviations (MADs) in marker-specific summary statistics (mean, standard deviation, and sparsity, defined as the percentage of zeros) between the SyNG-BTS augmented data and the empirical data are calculated as the pilot data sample size increases from 20 to 100. The MAD values are color-coded, with extremely large values represented as “>6”. Smaller MADs indicate better congruency between the augmented data and the empirical data. Each sub-panel column represents one of the three generative model families, and each row within a sub-panel corresponds to a specific model variant, as indicated on the left of each sub-panel. VAE1-10, MAF, and WGANGP consistently exhibit the smallest MAD values in their respective model families. (B) MADs in marker-specific statistics between the augmented data and the empirical data are evaluated when marker filtering is applied to pilot data. (C) Additional evaluation metrics, encompassing (i) the percentage of markers with non-zero counts in at least one sample (indicated as 1 – Pct(0-markers)), (ii) the agreement of sample clusters and data sources

SyNG-BTS successfully augmented two-group miRNA data

For the two-group setting, we replaced VAEs with CVAEs and excluded the GAN models due to their poor performance in the one-group setting. Detailed results are presented below for the TCGA SKCM/LAML study with marker filtering (Fig. 3; Fig. S5). Similar observations were noted for this study without marker filtering (Fig. S6) and for the BRCA/PRAD study (Fig. S7).

The performance of SyNG-BTS remained consistently strong in the two-group setting (Fig. 3; Fig. S5). Specifically, MAF once again emerged as the top performer, closely followed by CVAE1-10 (Fig. 3A and B; Fig. S5A). Their performance was influenced by pilot data sample size (Fig. 3A and B; Fig. S5A), marker filtering (Fig. 3 and Fig. S5 versus Fig. S6), offline augmentation (Fig. S5B), and hyperparameter tuning (Fig. S5C and D), similar to the one-group setting. Additionally, MAF consistently yielded superior results in terms of differential expression analysis, as indicated by the concordance correlation coefficients of *P*-values (Fig. 3A, third row) and group mean differences (Fig. 3A, fourth row). Notably, depth normalization, particularly with total count or TMM, proved to be more influential than in the one-group setting (Fig. 3B). It played a noticeable role in facilitating model training, especially for CVAE1-10, particularly for the inter-marker correlation metric and the two metrics related to differential expression analysis. The uniform manifold approximation and projection (UMAP) plot further affirmed the quality of the generated samples, displaying distinct separation by sample types without differentiation according to data sources, even with the runner-up generative model CVAE1-10 (Fig. 3C).

SyNG-BTS successfully augmented RNA data

For RNA-seq data augmentation, we focused on the better performing model variant for each DGM model based on the miRNA results, namely VAE, MAF, and WGANGP. Considering the substantial number of markers (60660) in RNA-seq data, we adjusted the loss ratio of VAE and CVAE to 1:100 (shorthand as VAE1-100 and CVAE1-100, respectively) and expanded the range of pilot data sample sizes to 50–250. Moreover, we excluded markers with both low mean and low variability across samples, reducing the number of markers to 1099 for the TCGA BRCA data and 1279 for the BRCA/PRAD data (see details in Supplementary Methods Table S1).

In the one-group setting, the performance of SyNG-BTS for RNA-seq aligned well with that for miRNA-seq (Fig. 4A–C; Fig. S8).

MAF performed the best for both marker-specific characteristics (Fig. 4A) and sample clustering (Fig. 4B), closely followed by VAE1-100. Like miRNA-seq, depth normalization had little impact on RNA-seq data augmentation in the one-group setting (Fig. 4C).

In the two-group setting, the performance of SyNG-BTS was again consistent with that for miRNA-seq (Fig. 4D–G; Fig. S9). MAF initially exhibited inferior performance to CVAE when the pilot data sample size was 50 but significantly improved as the sample size increased toward 250 (Fig. 4D and E). In particular, when the pilot data size exceeded 50, MAF demonstrated exceptional effectiveness in identifying differentially expressed markers between two sample types, achieving nearly perfect agreement with the empirical data in terms of the *P*-values and fold-changes (Fig. 4E). Both models were highly effective in sample clustering, with the identified clusters showing strong alignment with sample groups rather than data sources (Fig. 4E–G). The impact of depth normalization is mixed, with total-count and TMM facilitating smoother improvement over pilot data sample size for MAF (Fig. 4F).

For offline augmentation, the AE reconstruction approach faced challenges due to its complexity and the need for a relatively moderate marker-to-sample-size ratio in the pilot data (results not shown), while Gaussian noise addition proved to be more effective (Fig. 4B and E). We used the latter for RNA-seq data offline augmentation by combining an initial pilot dataset with nine noise-added datasets created by introducing Gaussian noise (see details in Supplementary Methods). This approach reduced variability in all evaluation metrics, thereby improving the quality of the augmented data, especially in the two-group setting (Fig. 4E). Unsurprisingly, the influence of offline augmentation was particularly marked when dealing with small pilot data sizes, with MAF reaping significant benefits in such instances.

Transfer learning enhanced the performance of SyNG-BTS

To examine the potential of transfer learning as a pretraining strategy for improving the performance of generative models, we pretrained VAEs with a loss ratio of 1:10 for miRNA-seq and 1:100 for RNA-seq, using datasets from one TCGA study or multiple studies combined (called the pretraining dataset) [70, 71]. The trained models were then used for model training with pilot datasets drawn from a different and intended TCGA study. As shown in Fig. 5, model training saw enhancement across all evaluated pilot data sizes. For miRNA-seq, the enhancement was particularly evident in the improvement of inter-marker relationship

when clustering a combined dataset of both generated and real samples, measured by the complementary adjusted Rand index (cARI), and (iii) the degree of correlation among member microRNAs belonging to the same polycistronic clusters, quantified by the concordance correlation coefficient of partial correlation coefficients (CCCPCC), are calculated across various pilot data sample sizes. Proximity of values for 1 – Pct(0-markers) to its level in the empirical data (indicated with a horizontal dashed line), along with elevated values of cARI and CCCPCC, signify improved congruency between the augmented data and the empirical data. Flow-based models exhibit smaller nonzero marker proportions than the empirical data, as they are above the dashed line; VAEs tend to generate approximately 50% of markers with zero counts in all samples, while GANs show the highest proportion of zero-count markers. In general, VAE and flow-based models outperform GAN models, with VAE1-10, MAF, and WGANGP emerging as the top performers in their respective model families. (D) The same additional evaluation metrics, including 1 – Pct(0-markers), cARI, and CCCPCC, are computed when marker filtering is applied to pilot data. (E) Evaluation metrics, including 1 – Pct(0-markers), cARI, and CCCPCC, are presented for the best performing variant in each generative model family, using two different training batch sizes (indicated by colors). VAE1-10 tends to be most sensitive to batch size, showing better performance for smaller batch sizes (i.e., deep training), while MAF and WGANGP tend to be insensitive. (F) Evaluation metrics, including 1 – Pct(0-markers), cARI, and CCCPCC, are presented for the best performing variant in each generative model family, using two different epoch strategies (indicated by colors). VAE1-10 prefers fixed epochs, while MAF already performs well with early stopping. (G) Evaluation metrics, including 1 – Pct(0-markers), cARI, and CCCPCC, are presented for the best performing variant in each generative model family, using three different depth normalization methods (indicated by colors): total count (TC), trimmed mean of M-values (TMM), and upper quartile (UQ), in comparison with no normalization (None) for pilot data. It is noteworthy that depth normalization has minimum impact on the generative model performance in this context. (H) Evaluation metrics, including 1 – Pct(0-markers), cARI, and CCCPCC, are presented with or without the use of offline augmentation via AE head (indicated by colors). It is evident that offline augmentation consistently improves the performance of all three generative models across evaluation metrics in terms of both the average value and variability. Unless stated otherwise, Panels A–G employ no offline augmentation, no depth normalization, a 10% batch fraction, a fixed epoch strategy for VAEs and GANs, and an early stopping strategy for flow-based models.

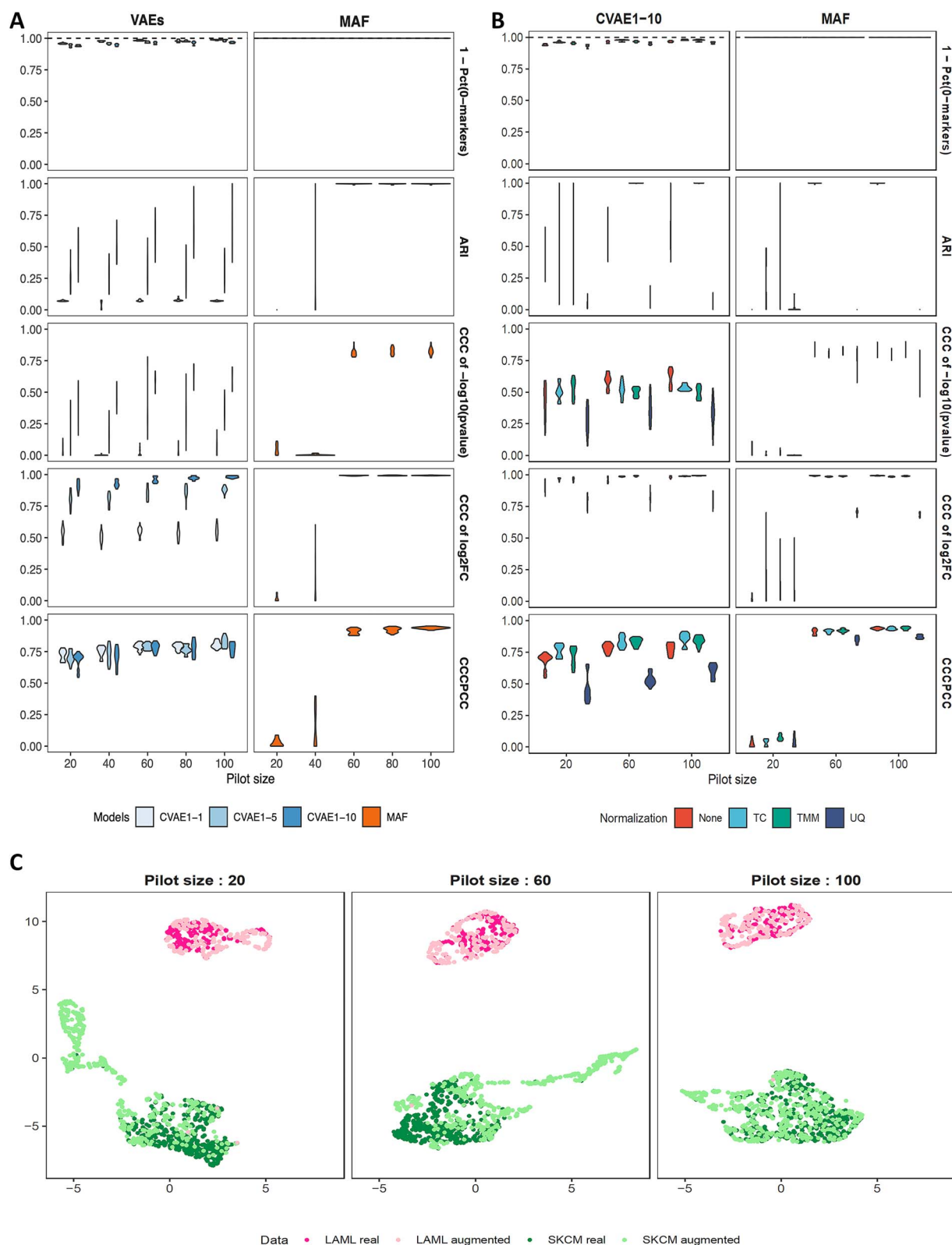


Figure 3. SyNG-BTS evaluation for microRNA-seq in the two-group setting, using pilot data from the combination of the TCGA SKCM and LAML studies with marker filtering. (A) Evaluation metrics assessing the congruence between the augmented data and the empirical data, including (i) 1 - Pct(0-markers); (ii) the agreement of sample clusters and sample types when clustering a combined dataset of both generated and real samples, measured by the adjusted Rand index (indicated as ARI); (iii) concordant correlation coefficient of P-values from differential expression analysis on the $-\log_{10}$ scale [indicated as CCC of $-\log_{10}(\text{P-value})$]; (iv) concordant correlation coefficient of \log_2 fold change from differential expression analysis (indicated as CCC of $\log_2\text{FC}$); and (v) CCCPC are calculated for various generative models as the pilot data sample size increases from 20 to 100 per sample group. (B) The same evaluation metrics for data congruence are calculated using three different depth normalization methods (indicated by colors) in comparison with no normalization. (C) The UMAP representation for the generated samples (by CVAE1-10) and the real samples, with the data source and the sample type indicated by colors.

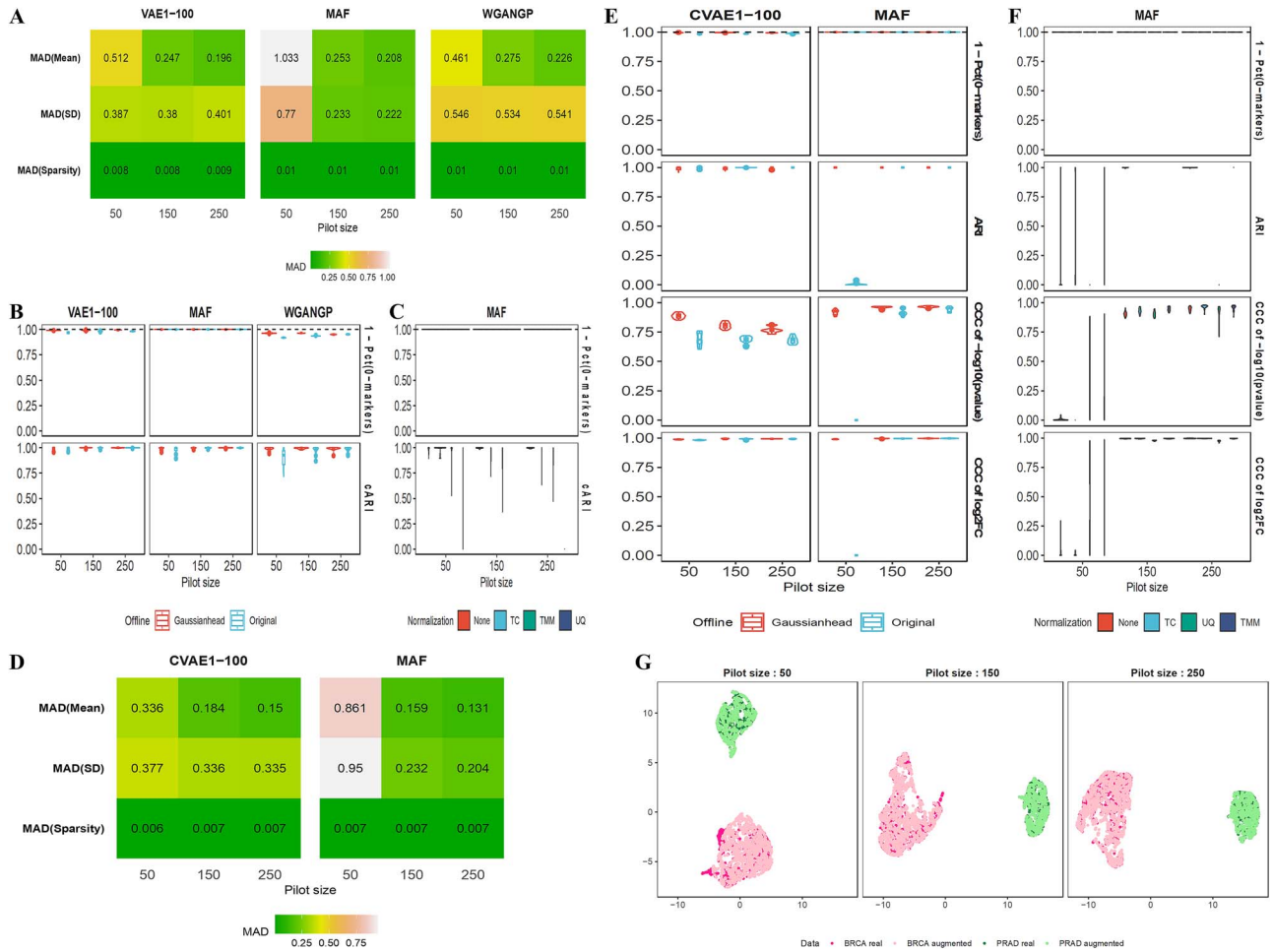


Figure 4. SyNG-BTS evaluation for RNA-seq in the one-group setting (Panels A–C), using pilot data from the TCGA BRCA study, and in the two-group setting (Panels D–G), using pilot data from the combination of the TCGA BRCA and PRAD studies, both with marker filtering. (A) MADs in marker-specific summary statistics (mean, standard deviation, and sparsity) between the augmented data and the empirical data are calculated as the pilot data sample size increases from 50 to 250. (B) Additional evaluation metrics for data congruence, including 1 - Pct(0-markers) and cARI, are calculated over varying pilot data sample sizes, with or without offline augmentation via Gaussian noise addition (indicated by colors). (C) Evaluation metrics for data congruence, including 1 - Pct(0-markers) and cARI, are calculated using three different depth normalization methods for pilot data (indicated by colors) in comparison with no normalization. (D) MADs in marker-specific summary statistics between the augmented data and the empirical data, are calculated as the pilot data sample size increases from 50 to 250 per sample group. (E) Evaluation metrics for data congruence, including 1 - Pct(0-markers) and ARI, are calculated with or without the use of offline augmentation via Gaussian noise addition (indicated by colors). (F) Evaluation metrics for data congruence are calculated using three different depth normalization methods for pilot data (indicated by colors) in comparison with no normalization. (G) The UMAP representation for the generated samples (by CVAE1-100) and the real samples for varying pilot data sample sizes, with the data source and the sample type indicated by colors.

(Fig. 5A, third row). Conversely, for RNA-seq, the enhancement was more remarkable in preserving the proportion of expressed markers (Fig. 5B, first row). While technically any dataset with the same set of markers can be used for pretraining, our findings highlighted the importance of the pretraining data having characteristics comparable to the pilot data (Fig. 5A, left column). Additionally, a larger pretraining dataset, such as the combination of TCGA PRAD, LAML, and SKCM data, led to greater enhancements compared to using the TCGA PRAD data alone, when augmenting pilot datasets drawn from the TCGA BRCA study (Fig. 5A, right column). These results underscored the value of incorporating transfer learning in transcriptomic data augmentation to leverage distributionally comparable and well-sized pretraining data.

SyntheSize successfully determined the sample size for miRNA studies

For demonstration purposes, we applied the SyntheSize approach for (post-hoc) sample size evaluation using the TCGA BRCA miRNA-seq data, which includes two subtypes: IDC and ILC

(Figs 1D and 6A). A subset of the TCGA BRCA miRNA-seq data (100 samples per subtype) was reserved as an independent validation set, utilized to provide a *de facto* assessment of the relationship between prediction accuracy and sample size. The remaining samples were then used as the input pilot data for SyntheSize algorithm. We computed accuracies for three learning techniques – support vector machine, k-nearest neighbors, and XGBoost – in classifying the two BRCA subtypes. The estimated accuracies fitted well with an IPLF curve, which began to plateau when the sample size reached about 50 per subtype, suggesting limited value in adding more samples (Fig. 6A, right column). Additionally, we obtained datasets with varying sample sizes by subsampling the validation set (up to 100 samples per subtype), assessed classification accuracies in these datasets, and fitted IPLF curves for the same three learning techniques (Fig. 6A, left column). The curves fitted to the empirical datasets closely mirrored those based on the augmented datasets, providing additional validation for the effectiveness of our proposed approach for sample size determination.

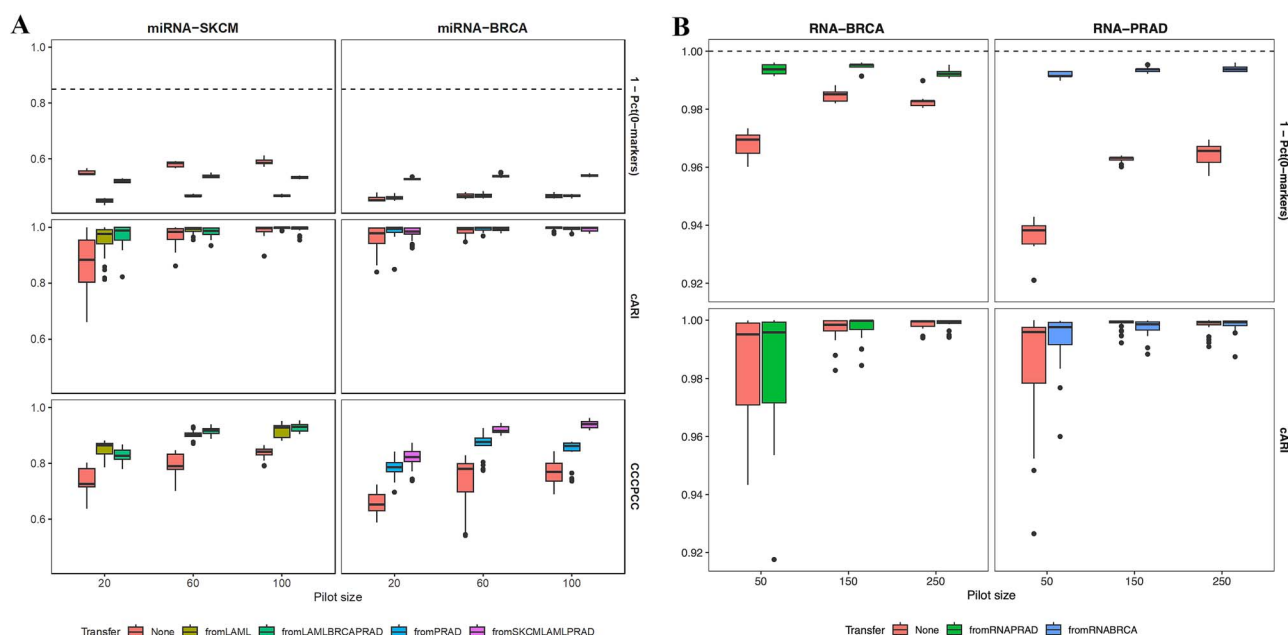


Figure 5. Evaluation of transfer learning for enhancing model training in SyNG-BTS using microRNA-seq data (Panel A) and RNA-seq data (Panel B) with marker filtering. (A) Evaluation metrics on the congruence of the augmented data and the empirical data, including $1 - \text{Pct}(0\text{-markers})$, cARI, and CCCPCC, are calculated when pilot data are drawn from the TCGA SKCM microRNA-seq study and models are pretrained using the TCGA LAML study or the combination of the TCGA BRCA, LAML, and PRAD studies (left column of sub-panels), and when pilot data are drawn from the TCGA BRCA microRNA-seq study and the models are pretrained using the TCGA PRAD study or the combination of the TCGA SKCM, LAML, and PRAD studies (right column of sub-panels). (B) Evaluation metrics for data congruence, including $1 - \text{Pct}(0\text{-markers})$ and cARI, are calculated when pilot data are drawn from the TCGA BRCA RNA-seq study and the models are pretrained using the TCGA PRAD study (left column of sub-panels), and when pilot data are drawn from the TCGA PRAD RNA-seq study and the models are pretrained using the TCGA BRCA study (right column of sub-panels).

SyntheSize successfully determined the sample size for RNA studies

Subsequently, we assessed SyntheSize using the TCGA BRCA RNA-seq data, similar to the assessment with the miRNA-seq data (Fig. 6B). Although exhibiting slightly inferior performance compared to its efficacy for miRNA-seq, SyntheSize provided a satisfactory sample size estimation for RNA-seq. This is evident from the proximity observed between the predicted accuracies using the augmented data and that derived from the empirical validation data.

For further illustration, we showcased SyntheSize in determining the sample size needed for building a predictor of immunotherapy response (CR/PR versus PD/SD), sourcing pilot RNA-seq data from a recent clinical study involving a PD-1 inhibitor, nivolumab, in patients with advanced clear cell renal cell carcinoma [61]. The real and generated samples had a high degree of similarity as revealed in the UMAP (Fig. S10). The accuracies of the three learning techniques again closely aligned with the IPLF model (Fig. 6C). The curves plateaued, indicating that their near-optimal accuracies were achieved, when the sample size reached about 200 samples per response group. Among the three techniques, k-nearest neighbors exhibited a noticeably smoother fit to the IPLF curve, albeit with a much higher sensitivity to the sample size as its performance floundered at low sample sizes, compared to support vector machine and XGBoost (Fig. 6C).

Discussion

Our proposed SyntheSize approach adeptly estimates the required sample size for machine learning with bulk transcriptomic sequencing data, harnessing the power of DGMs via

the SyNG-BTS algorithm to augment available pilot data. The consistent and reliable performance of SyntheSize, demonstrated in both miRNA-seq and RNA-seq, highlights its versatility and effectiveness in informing experimental design for transcriptomics studies using machine learning.

For illustrative purposes, we evaluated sample size for the TCGA BRCA study using the method developed by Dobbin and Simon (2007) [20], which is implemented in the R package *MKpower*. The estimated sample size varied significantly based on the tolerance parameter, ranging from approximately 6000 samples for a 1% tolerance to just four samples for a 5% tolerance. These estimates diverged greatly from the empirical sample size observed in our data, underscoring the limitations of such methods. This comparison supports the necessity and advantage of our proposed SyntheSize method.

We acknowledge that obtaining pilot data of reasonable size and good quality can be challenging, but it is necessary to avoid making parametric distribution assumptions or relying on a substantial empirical dataset for subsampling. Ideally, users should source their own pilot data that mirrors real-world data characteristics in the intended biomedical problem context. For the large dataset to be collected, for which the sample size is assessed, obtaining a pilot dataset of 40 to 60 samples is worthwhile. If this is not possible, users can turn to publicly available data. For instance, the TCGA offers high-quality transcriptomic sequencing data for 30 plus cancer types, each with hundreds of samples.

Through a comprehensive evaluation of SyNG-BTS in diverse settings, we have demonstrated the successful training of generative models for bulk transcriptomic sequencing data. The efficacy of these models is influenced by various factors related to the pilot data, such as its sample size and marker number, as well as specifications for the generative models, including model choice, hyperparameter tuning, and the use of offline augmentation and

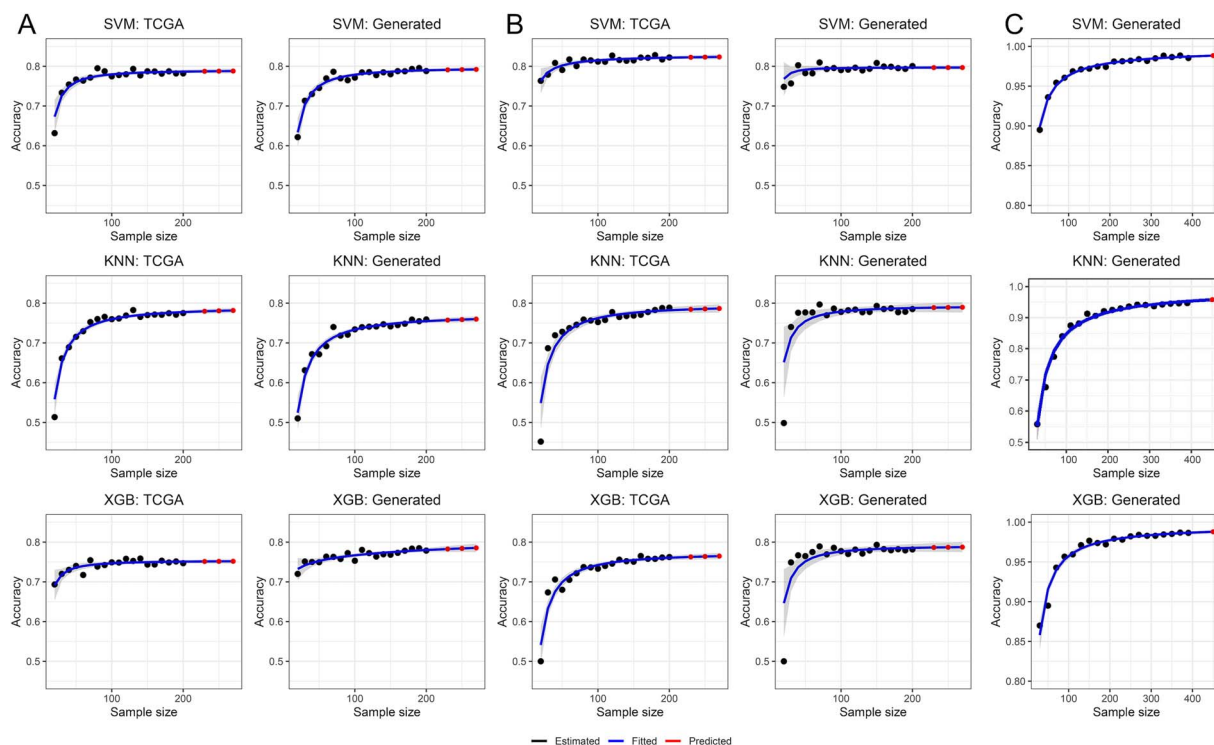


Figure 6. Evaluation of SyntheSize on microRNA-seq data (Panel A) and RNA-seq data (Panel B) from the TCGA BRCA study, and application of SyntheSize to RNA-seq data from a clinical study of nivolumab (Panel C). (A and B) Classifiers are constructed to distinguish the two breast cancer subtypes, IDC and ILC, in the TCGA BRCA study using empirical data (left column in each panel) or the SyNG-BTS augmented data (right column in each panel) and employing three machine learning techniques [top row in each panel: support vector machine (SVM); middle row: k-nearest neighbors (KNN); bottom row: XGBoost (XGB)]. (C) Classifiers are built to predict patient response to nivolumab, CR/PR and PD/SD, using RNA-seq data from a published clinical study as pilot data. In Panels A–C, classification accuracies are assessed for three learning techniques, including SVM, KNN, and XGB, across a range of sample sizes. Specifically, classification accuracies estimated from empirical or augmented data are plotted as black dots, while their fitted IPLFs are plotted as blue curves, projecting accuracies achieved at additional sample sizes indicated by red dots at the far right end of the fitted curves. The gray bands represent the 95% confidence regions for the fitted IPLFs.

transfer learning. Generally, model training is more successful when the pilot data maintains a reasonable marker-to-sample size ratio. In cases where this ratio is excessively high, the incorporation of offline augmentation and transfer learning has proven to be beneficial. The generative models need to be thoughtfully selected and meticulously tuned. Among the models investigated, MAF and VAE models consistently outperformed GAN models.

The runtime of the DGMs used in SyNG-BTS is an important consideration for its overall utility. In practice, the time required to train these models can vary based on data complexity, model architecture, and available computational resources. Our experiences found that the DGMs did not demand extensive computational resources, primarily due to the simplicity of the model structures employed and the modest size of the pilot datasets involved. Specifically, when using any of the DGMs in our studies, the runtime typically ranges between 1 and 5 minutes on a personal computer with 16GB of RAM and a 2.3 GHz Quad-Core Intel Core i5 processor. This brief runtime indicates the manageability of these models, affirming that even personal computers, without parallel computing setups, are sufficient for training and applying the DGMs. The low computational demands significantly broaden the potential for using SyntheSize to design transcriptomic sequencing studies, without necessitating high-end computing infrastructure.

In summary, our study demonstrated the successful training of generative models to augment bulk-tissue transcriptomic sequencing data, enabling effective sample size determination

using augmented datasets and the IPLF model. These computational resources are poised to greatly facilitate the deployment of supervised machine learning techniques in deriving effective sample classifiers from biomedical transcriptomic data. These contributions will significantly advance the development of essential computational tools crucial for designing classification studies with transcriptomic sequencing data, thereby accelerating their translation into clinically impactful predictors.

Key Points

- SyntheSize is a novel computational approach that establishes the accuracy-versus-sample size relationship for sample classification employing machine learning techniques.
- The performance of SyntheSize has been comprehensively assessed for microRNA and RNA sequencing data, considering diverse data characteristics and algorithm configurations.
- The Python and R code for implementing SyntheSize is available on GitHub.

Supplementary data

Supplementary data are available online at <https://academic.oup.com/bib>.

Acknowledgements

We thank Nicole Rusk for her insightful comments on the manuscript and Joseph Kanik for his assistance with the graphical design of Fig. 1.

Author contributions

Conceptualization: Y.Q. and L.X.Q.; Methodology: Y.Q., X.W., and L.X.Q.; Data Curation: Y.Q. and L.X.Q.; Formal Analysis: Y.Q., X.W., and L.X.Q.; Writing—Original Draft: Y.Q. and L.X.Q.; Writing—Review and Editing: Y.Q., X.W., and L.X.Q.; Resources: L.X.Q.; Supervision: L.X.Q.; Funding Acquisition: L.X.Q.

Funding

This work was supported by grants from the National Institutes of Health (HG012124 to Y.Q., X.W., and L.X.Q., CA214845 and CA008748 to L.X.Q.).

Data and code availability

The microRNA and RNA sequencing data used in this article are all publicly available, including those generated by the TCGA Research Network: <https://www.cancer.gov/tcga>. The Python and R code for implementing SyNG-BTS and SyntheSize is freely downloadable on GitHub (<https://github.com/LXQin/SyNG-BTS> and <https://github.com/LXQin/SyntheSize>). The R code for reproducing the results, which includes TCGA and immunotherapy study data downloading, data augmentation with SyNG-BTS, sample size assessment with SyntheSize, as well as the generation of all figures, can be found at <https://github.com/LXQin/SyntheSize-paper-supplementary-materials>.

References

- Van't Veer LJ, Bernards R. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature* 2008;**452**:564–70. <https://doi.org/10.1038/nature06915>.
- Adams JU. Genetics: big hopes for big data. *Nature* 2015;**527**:S108–9. <https://doi.org/10.1038/527S108a>.
- Lee AH. Prediction of cancer outcome with microarrays. *Lancet* 2005;**365**:1685. [https://doi.org/10.1016/S0140-6736\(05\)66541-5](https://doi.org/10.1016/S0140-6736(05)66541-5).
- Nair VS, Maeda LS, Ioannidis JP. Clinical outcome prediction by microRNAs in human cancer: a systematic review. *J Natl Cancer Inst* 2012;**104**:528–40. <https://doi.org/10.1093/jnci/djs027>.
- Pencina MJ, Peterson ED. Moving from clinical trials to precision medicine: the role for predictive modeling. *JAMA* 2016;**315**:1713–4. <https://doi.org/10.1001/jama.2016.4839>.
- Dhurandhar EJ, Vazquez AI, Argyropoulos GA. et al. Even modest prediction accuracy of genomic models can have large clinical utility. *Front Genet* 2014;**5**:417. <https://doi.org/10.3389/fgene.2014.00417>.
- Altman DG, Riley RD. Primer: an evidence-based approach to prognostic markers. *Nat Clin Pract Oncol* 2005;**2**:466–72. <https://doi.org/10.1038/ncponc0287>.
- Hey SP, Kesselheim AS. Countering imprecision in precision medicine. *Science* 2016;**353**:448–9. <https://doi.org/10.1126/science.aaf5101>.
- Hickey GL, Grant SW, Dunning J. et al. Statistical primer: sample size and power calculations—why, when and how? *Eur J Cardiothorac Surg* 2018;**54**:4–9. <https://doi.org/10.1093/ejcts/ezy169>.
- McKeigue P. Sample size requirements for learning to classify with high-dimensional biomarker panels. *Stat Methods Med Res* 2019;**28**:904–10. <https://doi.org/10.1177/0962280217738807>.
- Emanuel EJ, Wendler D, Grady C. What makes clinical research ethical? *JAMA* 2000;**283**:2701–11. <https://doi.org/10.1001/jama.283.20.2701>.
- Schulz KF, Grimes DA. Sample size slippages in randomised trials: exclusions and the lost and wayward. *Lancet* 2002;**359**:781–5. [https://doi.org/10.1016/S0140-6736\(02\)07882-0](https://doi.org/10.1016/S0140-6736(02)07882-0).
- Fang Z, Cui X. Design and validation issues in RNA-seq experiments. *Brief Bioinform* 2011;**12**:280–7. <https://doi.org/10.1093/bib/bbr004>.
- Li CI, Su PF, Shyr Y. Sample size calculation based on exact test for assessing differential expression analysis in RNA-seq data. *BMC Bioinformatics* 2013;**14**:357. <https://doi.org/10.1186/1471-2105-14-357>.
- Busby MA, Stewart C, Miller CA. et al. Scotty: a web tool for designing RNA-seq experiments to measure differential gene expression. *Bioinformatics* 2013;**29**:656–7. <https://doi.org/10.1093/bioinformatics/btt015>.
- Bi R, Liu P. Sample size calculation while controlling false discovery rate for differential expression analysis with RNA-sequencing experiments. *BMC Bioinformatics* 2016;**17**:146. <https://doi.org/10.1186/s12859-016-0994-9>.
- Wu H, Wang C, Wu Z. PROPER: comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics* 2015;**31**:233–41. <https://doi.org/10.1093/bioinformatics/btu640>.
- Yu L, Fernandez S, Brock G. Power analysis for RNA-Seq differential expression studies. *BMC Bioinformatics* 2017;**18**:234. <https://doi.org/10.1186/s12859-017-1648-2>.
- Vieth B, Ziegenhain C, Parekh S. et al. powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* 2017;**33**:3486–8. <https://doi.org/10.1093/bioinformatics/btx435>.
- Dobbin KK, Simon RM. Sample size planning for developing classifiers using high-dimensional DNA microarray data. *Biostatistics* 2007;**8**:101–17. <https://doi.org/10.1093/biostatistics/kxj036>.
- Dobbin KK, Zhao Y, Simon RM. How large a training set is needed to develop a classifier for microarray data? *Clin Cancer Res* 2008;**14**:108–14. <https://doi.org/10.1158/1078-0432.CCR-07-0443>.
- de Valpine P, Bitter HM, Brown MP. et al. A simulation-approximation approach to sample size planning for high-dimensional classification studies. *Biostatistics* 2009;**10**:424–35. <https://doi.org/10.1093/biostatistics/kxp001>.
- Mukherjee S, Tamayo P, Rogers S. et al. Estimating dataset size requirements for classifying DNA microarray data. *J Comput Biol* 2003;**10**:119–42. <https://doi.org/10.1089/106652703321825928>.
- Figueroa RL, Zeng-Treitler Q, Kandula S. et al. Predicting sample size required for classification performance. *BMC Med Inform Decis Mak* 2012;**12**:8. <https://doi.org/10.1186/1472-6947-12-8>.
- Friedman LM, Furberg CD, DeMets DL. et al. *Fundamentals of Clinical Trials*. New York, NY, USA: Springer, 2015.
- Li J, Lenferink AEG, Deng Y. et al. Identification of high-quality cancer prognostic markers and metastasis network modules. *Nat Commun* 2010;**1**:34. <https://doi.org/10.1038/ncomms1033>.
- Mobadersany P, Yousefi S, Armgad M. et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc Natl Acad Sci USA* 2018;**115**:E2970–9. <https://doi.org/10.1073/pnas.1717139115>.
- Ching T, Himmelstein DS, Beaulieu-Jones BK. et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* 2018;**15**:20170387. <https://doi.org/10.1098/rsif.2017.0387>.
- Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

30. Foster D. *Generative Deep Learning*. Sebastopol, CA, USA: O'Reilly Media, Inc., 2022.
31. Thabane L, Ma J, Chu R. et al. A tutorial on pilot studies: the what, why and how. *BMC Med Res Methodol* 2010;**10**:1. <https://doi.org/10.1186/1471-2288-10-1>.
32. Shrager J, Hogg T, Huberman BA. A graph-dynamic model of the power law of practice and the problem-solving fan-effect. *Science* 1988;**242**:414–6. <https://doi.org/10.1126/science.3175664>.
33. Cortes C, Jackel LD, Solla S. et al. Learning curves: asymptotic values and rate of convergence. *Adv Neural Inf Proces Syst* 1993;**6**.
34. Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: a review. *Med Image Anal* 2019;**58**:101552. <https://doi.org/10.1016/j.media.2019.101552>.
35. Marouf M, Machart P, Bansal V. et al. Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nat Commun* 2020;**11**:166. <https://doi.org/10.1038/s41467-019-14018-z>.
36. Treppner M, Salas-Bastos A, Hess M. et al. Synthetic single cell RNA sequencing data from small pilot studies using deep generative models. *Sci Rep* 2021;**11**:9403. <https://doi.org/10.1038/s41598-021-88875-4>.
37. Heydari AA, Davalos OA, Zhao L. et al. ACTIVA: realistic single-cell RNA-seq generation with automatic cell-type identification using introspective variational autoencoders. *Bioinformatics* 2022;**38**:2194–201. <https://doi.org/10.1093/bioinformatics/btac095>.
38. Rezende D, Mohamed S. Variational inference with normalizing flows. In: *International Conference on Machine Learning*. PMLR, 2015, 1530–8.
39. Goodfellow I. et al. Generative adversarial nets. *Adv Neural Inf Proces Syst* 2014;**27**.
40. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv:1312.6114. 2013, preprint: not peer-reviewed.
41. Dinh L, Krueger D, Bengio Y. Nice: non-linear independent components estimation. arXiv, arXiv:1410.8516. 2014, preprint: not peer-reviewed.
42. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004;**116**:281–97. [https://doi.org/10.1016/S0092-8674\(04\)00045-5](https://doi.org/10.1016/S0092-8674(04)00045-5).
43. Ambros V. The functions of animal microRNAs. *Nature* 2004;**431**:350–5. <https://doi.org/10.1038/nature02871>.
44. Davila Delgado JM, Oyedele L. Deep learning with small datasets: using autoencoders to address limited datasets in construction management. *Appl Soft Comput* 2021;**112**:107836. <https://doi.org/10.1016/j.asoc.2021.107836>.
45. Kramer MA. Nonlinear principal component analysis using autoassociative neural networks. *AICHE J* 1991;**37**:233–43. <https://doi.org/10.1002/aic.690370209>.
46. Bishop CM. Training with noise is equivalent to Tikhonov regularization. *Neural Comput* 1995;**7**:108–16. <https://doi.org/10.1162/neco.1995.7.1.108>.
47. Holmstrom L, Koistinen P. Using additive noise in back-propagation training. *IEEE Trans Neural Netw* 1992;**3**:24–38. <https://doi.org/10.1109/72.105415>.
48. Wickens CD, Helton WS, Hollands JG. et al. *Engineering Psychology and Human Performance*. Abingdon, Oxfordshire, UK: Routledge, 2021.
49. Network CGAR. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 2008;**455**:1061–8. <https://doi.org/10.1038/nature07385>.
50. Chu A, Robertson G, Brooks D. et al. Large-scale profiling of microRNAs for The Cancer Genome Atlas. *Nucleic Acids Res* 2016;**44**:e3. <https://doi.org/10.1093/nar/gkv808>.
51. Colaprico A, Silva TC, Olsen C. et al. TCGAAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* 2016;**44**:e71. <https://doi.org/10.1093/nar/gkv1507>.
52. Bullard JH, Purdom E, Hansen KD. et al. Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics* 2010;**11**:94. <https://doi.org/10.1186/1471-2105-11-94>.
53. Qin LX, Zou J, Shi J. et al. Statistical assessment of depth normalization for small RNA sequencing. *JCO Clin Cancer Inform* 2020;**4**:567–82. <https://doi.org/10.1200/CCI.19.00118>.
54. Zou J, Düren Y, Qin LX. PRECISION.Seq: an R package for benchmarking depth normalization in microRNA sequencing. *Front Genet* 2021;**12**:823431. <https://doi.org/10.3389/fgene.2021.823431>.
55. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res* 2003;**3**:1157–82.
56. Howell DC. Median absolute deviation. Wiley StatsRef 2014. <https://doi.org/10.1002/9781118445112.stat06232>.
57. Düren Y, Lederer J, Qin LX. Depth normalization of small RNA sequencing: using data and biology to select a suitable method. *Nucleic Acids Res* 2022;**50**:e56. <https://doi.org/10.1093/nar/gkac064>.
58. Lawrence I, Lin K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989;**45**:255. <https://doi.org/10.2307/2532051>.
59. Rand WM. Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc* 1971;**66**:846–50. <https://doi.org/10.1080/01621459.1971.10482356>.
60. McInnes L, Healy J, Melville J. Umap: uniform manifold approximation and projection for dimension reduction. arXiv, arXiv:1802.03426, 2018, preprint: not peer-reviewed.
61. Braun DA, Hou Y, Bakouny Z. et al. Interplay of somatic alterations and immune infiltration modulates response to PD-1 blockade in advanced clear cell renal cell carcinoma. *Nat Med* 2020;**26**:909–18. <https://doi.org/10.1038/s41591-020-0839-y>.
62. Bojanowski P, Joulin A, Lopez-Paz D. et al. Optimizing the latent space of generative networks. arXiv, arXiv:1707.05776, 2017, preprint: not peer-reviewed.
63. Arjovsky M, Chintala S, Bottou L. Wasserstein Generative Adversarial Networks, pp. 214–23. PMLR.
64. Gulrajani I, Ahmed F, Arjovsky M. et al. Improved training of Wasserstein GANs. *Adv Neural Inf Proces Syst* 2017;**30**.
65. Papamakarios G, Pavlakou T, Murray I. Masked autoregressive flow for density estimation. *Adv Neural Inf Proces Syst* 2017;**30**.
66. Dinh L, Sohl-Dickstein J, Bengio S. Density estimation using real NVP. arXiv, arXiv:1605.08803, 2016, preprint: not peer-reviewed.
67. Kingma DP, Dhariwal P. Glow: generative flow with invertible 1x1 convolutions. *Adv Neural Inf Proces Syst* 2018;**31**.
68. Brownlee J. A gentle introduction to early stopping to avoid overtraining neural networks. *Mach Learn Mastery* 2019. <https://machinelearningmastery.com/early-stopping-to-avoid-overtraining-neural-network-models/>. Accessed Oct 31, 2023.
69. Asperti A, Trentin M. Balancing reconstruction error and Kullback-Leibler divergence in variational autoencoders. *IEEE Access* 2020;**8**:199440–8. <https://doi.org/10.1109/ACCESS.2020.3034828>.
70. Pratt L, Jennings B. A survey of transfer between connectionist networks. *Connect Sci* 1996;**8**:163–84. <https://doi.org/10.1080/095400996116866>.
71. Wang J, Agarwal D, Huang M. et al. Data denoising with transfer learning in single-cell transcriptomics. *Nat Methods* 2019;**16**:875–8. <https://doi.org/10.1038/s41592-019-0537-1>.