Article

Patterns

GAiN: An integrative tool utilizing generative adversarial neural networks for augmented gene expression analysis

Highlights

- Genomics researchers are often faced with the problem of limited sample sizes
- Limited samples occur in rare disease, minority populations, and limited resources
- GAiN improves expression analysis with limited sample numbers
- GAiN's benefit is critical for downstream pathway analysis

Authors

Michael R. Waters, Matthew Inkman, Kay Jayachandran, ..., Obi L. Griffith, Jeffrey J. Szymanski, Jin Zhang

Correspondence

jin.zhang@wustl.edu

In brief

Genomic researchers are faced with the challenge of analyzing cohorts with limited samples. To address this need, the authors developed an integrative approach, GAiN, to capture patterns of gene expression from unique small datasets on the basis of an ensemble of generative adversarial networks while leveraging big population data. GAiN is reliable in discovering differentially expressed genes and enriched pathways where conventional methods are inadequate.

Waters et al., 2024, Patterns *5*, 100910
February 9, 2024
https://doi.org/10.1016/j.patter.2023.100910

CellPress

Patterns

CellPress
OPEN ACCESS

Article

GAiN: An integrative tool utilizing generative adversarial neural networks for augmented gene expression analysis

Michael R. Waters,[1] Matthew Inkman,[1] Kay Jayachandran,[1] Roman M. Kowalchuk,[2] Clifford Robinson,[1,3] Julie K. Schwarz,[1,3,4] S. Joshua Swamidass,[5,6,7] Obi L. Griffith,[8,9] Jeffrey J. Szymanski,[1,3] and Jin Zhang[1,3,10,11,*]

[1]Department of Radiation Oncology, Washington University School of Medicine, St. Louis, MO 63108, USA
[2]Department of Radiation Oncology, Mayo Clinic, Rochester, MN 55905, USA
[3]Alvin J. Siteman Cancer Center, Washington University School of Medicine, St. Louis, MO 63110, USA
[4]Department of Cell Biology and Physiology, Washington University School of Medicine, St. Louis, MO 63108, USA
[5]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63110, USA
[6]Department of Biomedical Engineering, Washington University in St. Louis, St. Louis, MO 63105, USA
[7]Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, MO 63105, USA
[8]Department of Medicine, Washington University School of Medicine, St. Louis, MO 63110, USA
[9]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO 63110, USA
[10]Institute for Informatics (I^2), Washington University School of Medicine, St. Louis, MO 63110, USA
[11]Lead contact
*Correspondence: jin.zhang@wustl.edu
https://doi.org/10.1016/j.patter.2023.100910

THE BIGGER PICTURE Studying the expression of genes within tissue samples is a common method of exploring the biology and behavior of that tissue. Comparing differences in gene expression between two biological groups can identify distinctions that drive biological behavior. In the context of medicine, such analysis can reveal biomarkers for different disease states and even suggest treatment targets. Such differential gene expression analysis, however, can lack both sensitivity and accuracy when low numbers of samples are available for RNA sequencing. The tool presented in this paper, GAiN, can enhance the ability of researchers to accurately identify true-positive gene expression differences and pathways between phenotypic groups by leveraging machine learning to uncover the structural gene expression patterns of even small numbers of biological samples.

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

Big genomic data and artificial intelligence (AI) are ushering in an era of precision medicine, providing opportunities to study previously under-represented subtypes and rare diseases rather than categorize them as variances. However, clinical researchers face challenges in accessing such novel technologies as well as reliable methods to study small datasets or subcohorts with unique phenotypes. To address this need, we developed an integrative approach, GAiN, to capture patterns of gene expression from small datasets on the basis of an ensemble of generative adversarial networks (GANs) while leveraging big population data. Where conventional biostatistical methods fail, GAiN reliably discovers differentially expressed genes (DEGs) and enriched pathways between two cohorts with limited numbers of samples (n = 10) when benchmarked against a gold standard. GAiN is freely available at GitHub. Thus, GAiN may serve as a crucial tool for gene expression analysis in scenarios with limited samples, as in the context of rare diseases, under-represented populations, or limited investigator resources.

INTRODUCTION

The combination of big genomic data and novel artificial intelligence (AI) technologies is ushering in an era of precision medicine, providing new opportunities for and posing new challenges to the study of previously under- or un-represented subtypes and rare diseases.[1] The majority of clinical researchers can still only rely on conventional biostatistical methods to study small datasets with unique phenotypes that may or may not be represented in large public datasets. Alarmingly, low sample numbers and small effect sizes have recently been identified as leading threats to research validity and reproducibility.[2,3] Among the most fundamental and widely applicable research needs are gene expression and pathway analyses for clinically relevant biological phenotypes.[4,5] However, gene expression patterns are complex and governed via nonlinear interactions of thousands of gene products.[6–8] This complexity stands as a major hurdle to generating robust mechanistic conclusions using existing methods when samples are limited.[9,10] To put this in perspective, using the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database as an example, the overwhelming majority (approximately 75%) of previously published datasets include fewer than 20 samples per dataset (Figure S1).[11] Given this, the aggregate impact of improving gene expression and pathways analysis in the context of small sample cohorts is extremely significant.

In this study, we apply cutting-edge AI technologies to address the challenges of robust gene expression pattern recognition from a small dataset, while also leveraging the existing resource of big public genomics data. The burgeoning field of data augmentation using deep learning techniques has shown the ability to increase the number and diversity of observations while preserving the underlying data structure, including nonlinear relationships between data features.[12] A recently developed generative modeling technique, generative adversarial networks (GANs), has achieved remarkable results in the fields of computer vision, natural language processing, and medical image analysis.[12,13]

Considering this, we developed GAiN, a tool leveraging GANs and existing public data to augment gene expression analysis of datasets with limited sample numbers.[14] To our knowledge, this is the first tool which uses GAN deep learning data augmentation for supervised downstream differential expression (DE) and pathway analysis. We established the ability of GAiN, which includes an ensemble of 20 GANs and a random forest model, trained on small numbers (i.e., n = 10) of samples for each phenotype, to discover gene expression differences and enriched pathways which one would observe if they collected hundreds or thousands of biological samples. To test this, we downloaded >4,000 samples from The Cancer Genome Atlas (TCGA), covering 4 large cohorts of samples to establish gold-standard benchmarks for subtype phenotypes.[15] Our benchmarking results showed that, while traditional biostatistical methods struggled to robustly predict patterns of differential gene expression demonstrated by analysis of a gold-standard comparison when applied to cohorts of small samples, GAiN reliably discovered differentially expressed genes (DEGs) using only small datasets. Furthermore, we show in a variety of scenarios that downstream pathway enrichment analysis (PEA) is greatly enhanced by using GAiN.

As GAN models are intrinsically hard to train, we provide an automated tool for researchers studying any disease or phenotype. Our tool is very efficient, and can be executed on a standard laptop, with a typical runtime of several hours. The GAiN tool is free and open source software and can be downloaded at https://github.com/jin-wash-u/GAiN.

RESULTS

Overview of the GAiN tool

The goal of the GAiN tool is to generate a large synthetic population of gene expression profiles that model a particular biological phenotype by using GANs trained on a small number of samples and fine-tuned using existing large, public genomic data. Next, these large cohorts of synthetic gene expression profiles are used for differential gene expression testing, with the results available for downstream PEA and additional genomic exploration. We hypothesize that GAiN augmented analysis substantially improves upon classical biostatistical techniques, especially in the context of limited samples. To this end, we implemented our data augmentation tool, GAiN, and trained it on small numbers of samples from two separate phenotypes of interest. GAiN is designed to train an ensemble of GANs (a customizable hyperparameter; n = 5 by default) on normalized expression data from two phenotypes of interest (Figure 1A; also see experimental procedures). Each network consists of two sub-models, a generative model (i.e., a denoising autoencoder [DAE]) and a discriminative model (i.e., a multilayer perceptron), for each phenotypic group (Figure 1B). GAiN uses the generators trained on the real samples to generate a large number (500 by default) of Z-normalized synthetic samples for the corresponding phenotype, and restores scale to this expression data using population data (Figure 1C; also see experimental procedures). Last, downstream DE testing (via edgeR) and PEA (via enrichR) is performed between the two GAiN generated cohorts to identify gene expression patterns and enriched pathways which one would expect to observe only if they collected hundreds or thousands of biological samples using traditional gene expression analysis methods[16,17] (Figure 1D). The GAiN augmentation tool for integrative and supervised gene expression analysis is freely available at https://github.com/jin-wash-u/GAiN.

Comparison with gold-standard benchmarks

Although GAiN can be applied to any dataset, here we used TCGA data to build reliable gold-standard benchmarks leveraging TCGA's large sample sizes and well-defined cancer phenotypes. Altogether, we downloaded >4,000 samples covering 4 large cancer cohorts in order to benchmark DEGs and altered pathways among 10 cancer-related phenotypes (Table S1). For each analysis, 50% of all samples were randomly reserved as a population cohort (PC), with the other 50% as a test cohort (TC) (Figure 1D). Gold-standard benchmarks for DEGs and enriched pathways were generated comparing the two phenotypes of the TC data, using several hundred samples per phenotype in each comparison. Additionally, for each phenotype, 10 samples (denoted as "small" or "sparse" in this study) from the PC were provided to GAiN to generate an augmented dataset. The two augmented datasets were then analyzed for

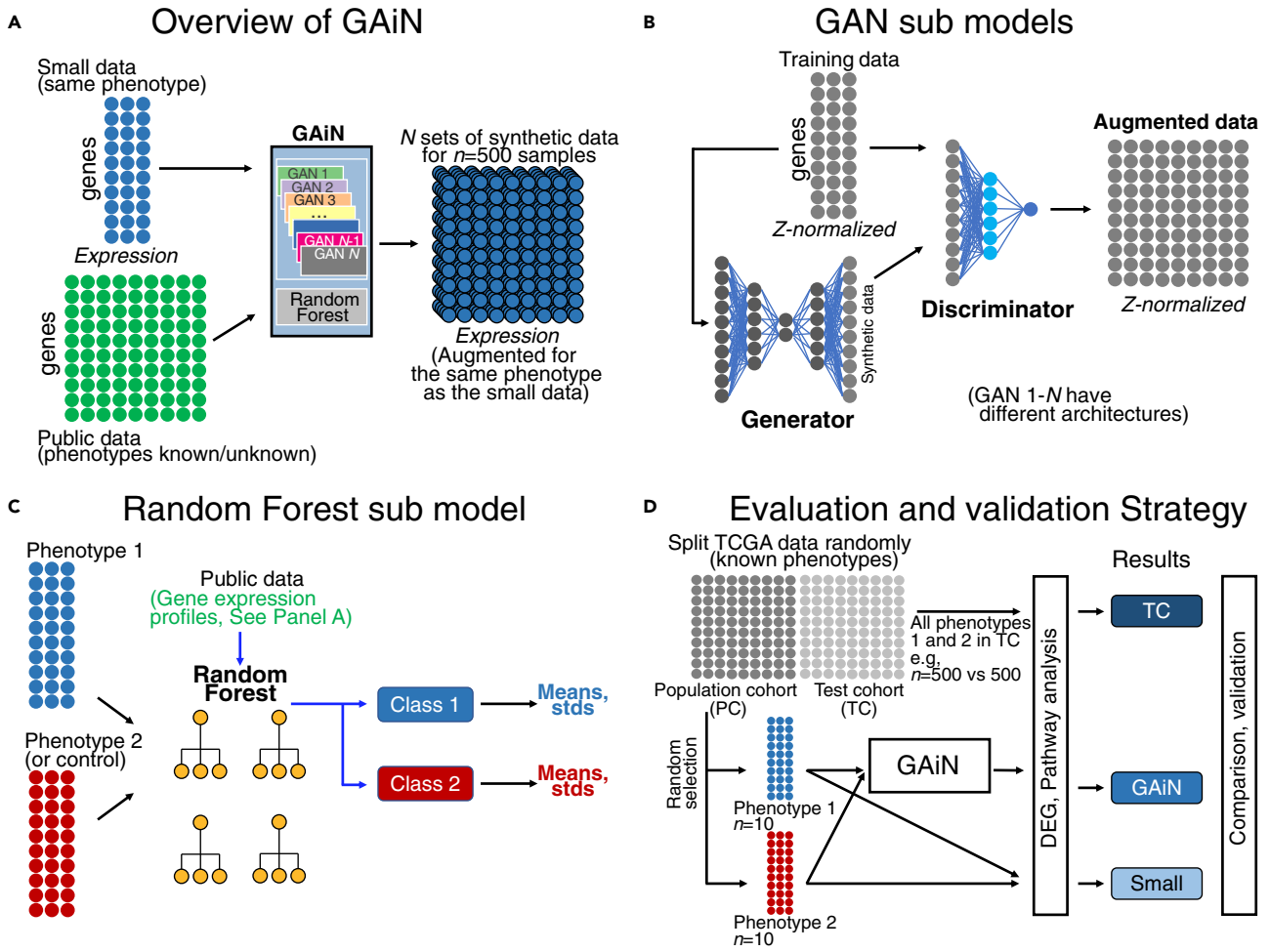


Figure 1. Overview of the GAiN tool

(A) GAiN accepts normalized gene expression data from a number of samples (typically small) of two phenotypes of interest and from a large population cohort chosen from public data that matches the biology of the phenotypes of interest as closely as possible (it may contain samples of other phenotypes). From these inputs, GAiN uses an ensemble of generative adversarial networks (GANs) and a random forest classifier to produce a series of gene expression datasets with large numbers of synthetic samples that encode the gene expression of each phenotype.

(B) Each GAN within the ensemble consists of two sub-models: a generator that learns from *Z*-normalized gene expression data of a single phenotype and attempts to generate synthetic outputs indistinguishable from its training data; and a discriminator that accepts both real training data and the synthetic data from the generator as input and attempts to correctly label each sample as real or synthetic. The generator and discriminator are trained in alternating fashion for 500 epochs, at which point the generator output is indistinguishable from the real training data and a large set of synthetic samples of the desired phenotype is output.

(C) The gene expression data for the synthetic samples output by the GAN is *Z*-normalized, so in order to restore scale to the individual genes, GAiN uses normalized gene expression values from public data. A random forest classifier is trained on the real expression data from each of the two phenotypes, then used to label each sample in the public data as more similar to phenotype 1 or phenotype 2. The means and SDs of each gene in the public samples assigned to each phenotype are then used to restore absolute gene expression values to the *Z*-normalized synthetic data.

(D) In order to evaluate GAiN's ability to reveal true differences in gene expression, large gene expression cohorts for pairs of phenotypes of interest were obtained from TCGA. These samples were stratified by phenotype and randomly divided into a population cohort (PC) and a test cohort (TC). DE gene analysis between the phenotypes in the TC provided gold-standard DE gene and enriched pathway lists against which to benchmark GAiN's results. A small cohort of 20 samples, 10 from each phenotype, was randomly selected from the PC, and the small cohort was submitted to GAiN as training data with the full PC as population data. DE gene and enriched pathway results from GAiN and from the un-augmented small cohort were then compared with those of the TC.

DEGs and enriched pathways, and the results were compared with the gold-standard benchmarks. To measure improvements gained by the GAN data augmentation approach, we compared the results from differential gene expression analysis and pathway analysis using the gold-standard TC, the GAiN augmented analysis, and the un-augmented small datasets. Results from gold-standard TC data are labeled "TC," results from the un-augmented small data are labeled "small," and results from the GAiN augmented data are labeled "GAiN" (Figure 1D).

Augmentation of gene expression analysis using small sample numbers

We assessed whether GAiN augmentation of two cohorts of limited sample size, and downstream DE testing between the

two augmented cohorts, improves DE analysis compared with using limited sample numbers alone. To test this hypothesis, we initially evaluated GAiN augmentation of two phenotypes from the same disease site: we downloaded TCGA gene expression data of lung adenocarcinoma (LUAD; n = 533) and lung squamous cell carcinoma (LUSC; n = 502) to serve as our two cohorts. As stated previously, 50% of the LUAD and LUSC cohort was randomly reserved as a PC, with the other 50% as a TC. The samples for each "small" dataset (n = 10 each phenotype) were randomly selected from the PC. Thus, we provided GAiN with only ~2% of the original TCGA-LUAD (n = 10 of 533) and TCGA-LUSC (n = 10 of 502) data. Five hundred synthetic LUAD and 500 synthetic LUSC samples were generated using these sparse samples by training 20 GANs in the GAiN tool. Given the large size of the synthetic and TC datasets, a large fraction of the total genes tested were deemed significant by multiple hypothesis testing. For useful discrimination, we used a rank-based method and compared the top 1,000 DEGs, as determined by edgeR false discovery rate (FDR)-adjusted p value, from DE testing between the GAiN synthetic cohorts, the TCs, and the non-augmented small samples (Data S1). First, we observed that the synthetic gene expression profiles generated by GAiN, using only sparse sets from the PC data, closely mirrored the expression profiles of both cancer types of genes in the TC data (Data S2). Principal-component analysis (PCA) of samples from the TC and GAiN cohorts (Figure S2) and Bhattacharyya distances calculated between the resulting distributions (Table S2) confirm the GAiN-generated samples from each phenotype have a very similar expression profile to the corresponding TC samples. For illustration, Figures 2A and 2B demonstrate GAiN augmented and TC expression values for the top 3 most and least DEGs by p value in the TC data. In all cases, synthetic data captures not only whether these genes are reliably expressed in the phenotype of interest, but also reliably mirrors real data expression levels. This also holds true for expression of genes known to be markers for biological differences between adeno and squamous cell carcinomas (Figure S3). Next, we found that the top 1,000 genes identified from DE analysis of the GAiN augmented datasets rediscovered 800 of 1,000 of the true-positive DEGs from the TC (F1 = 0.8; Figures 2C, 3, and S4). By comparison, the un-augmented small dataset predicted 602 of 1,000 genes (F1 = 0.6), and F1 score fell farther when only using genes significant by multiple hypothesis using edgeR, as is the approach in conventional biostatistical analysis (F1 = 0.56; Figures 2C and 3). The strong performance of GAiN synthetic samples in recapitulating true phenotypic differences in patterns of gene expression from the TC across different DEG thresholds (comparing the top n genes; e.g., n = 1,000 in Figure 2C) was demonstrated using receiver operating characteristic (ROC) curve analysis (Figure S5, lung), in which GAiN achieved an area under the curve (AUC) of 0.91. This performance is further illustrated by rank-rank comparison of genes ranked by DE between phenotypes in the TC and GAiN cohorts, with a high correlation of 0.89 between the lists (Figures S6 and S7).

GAiN applicability to a broad range of phenotypes

Using limited sample numbers and GAiN augmentation, we found that we were able to largely reproduce a DEG list one would observe if they collected thousands of biological samples in a comparison of LUAD and LUSC. Additionally, we found that using GAiN substantially improves upon a conventional biostatistical approach to analyzing a limited number of samples. We next assessed the generalizability of the GAiN workflow by repeating a comparison of DEGs from analysis of TC data, GAiN augmented data, and small data in other biological scenarios. First, GAiN performed similarly well, with improved F1 in discovery of gold-standard DEGs over a comparison of non-augmented small data from TCGA datasets of kidney renal papillary carcinoma (TCGA-KIRP), and kidney clear cell carcinoma (TCGA-KIRC), in which GAiN achieved an F1 score of 0.72 for the top 1,000 genes. In comparison, using an edgeR workflow on the un-augmented small dataset, the F1 score for re-discovering the gold-standard TC DEGs was 0.64 (Figure 3). When the gene threshold is varied, GAiN achieves AUC = 0.89 (Figure S5). We next assessed GAiN's accuracy in detecting gene expression differences between cancers of similar histology but different pathological stage by comparing low-grade glioma (TCGA-LGG) and glioblastoma (TCGA-GBM). Although GAiN achieved an F1 score of 0.72, edgeR only had F1 = 0.64 when comparing the top 1,000 genes with the TC (Figure 3); varying the gene threshold, AUC = 0.92 (Figure S5). Furthermore, we tested GAiN's workflow when comparing tumors of differing molecular subtypes by comparing triple negative and luminal B breast cancer subtypes from TCGA breast invasive carcinoma (TCGA-BRCA) dataset (Figure 3). The F1 scores for GAiN and edgeR comparing 1,000 genes with the TC are 0.75 and 0.57, respectively (Figure 3); varying the threshold, GAiN's AUC = 0.89 (Figure S5). Importantly, we demonstrated that GAiN is effective when using a heterogeneous PC by repeating the breast cancer analysis using the entire TCGA-BRCA as the PC (containing luminal A, luminal B, HER2, triple-negative, and adjacent normal breast samples; Figure S8). In total, these results demonstrated GAiN's generalizable ability to augment datasets of small sample size for improved DE analysis.

Augmented data for accurate pathway analysis

To assess the biological relevance of GAiN augmentation by assessing pathway dynamics between phenotypes, we performed PEA using the Gene Ontology (GO) pathway database on the foregoing DEG results from GAiN and edgeR comparing 10 TCGA-LUAD and 10 TCGA-LUSC samples.[18] To generate a gold-standard list of enriched pathways, we performed DE analysis and enrichR PEA between LUAD and LUSC in the TC. We next performed PEA on DEGs obtained through GAiN analysis and edgeR analysis of the small sample set and compared the top enriched pathways for the three workflows. Strikingly, 7 of the top 10 enriched GO pathways from the GAiN gene list matched the gold-standard list, while only 2 of 10 of the PEA results from the non-augmented DEGs matched (Figure 4). Similar improvements were observed when analyzing the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome databases and when analyzing all statistically enriched pathways as opposed to the top 10 pathways[19,20] (Figure S9; Data S3–S5). Given these findings, we concluded that using GAiN greatly improves the ability of pathway level analysis to discover the most relevant biological signals when using limited sample

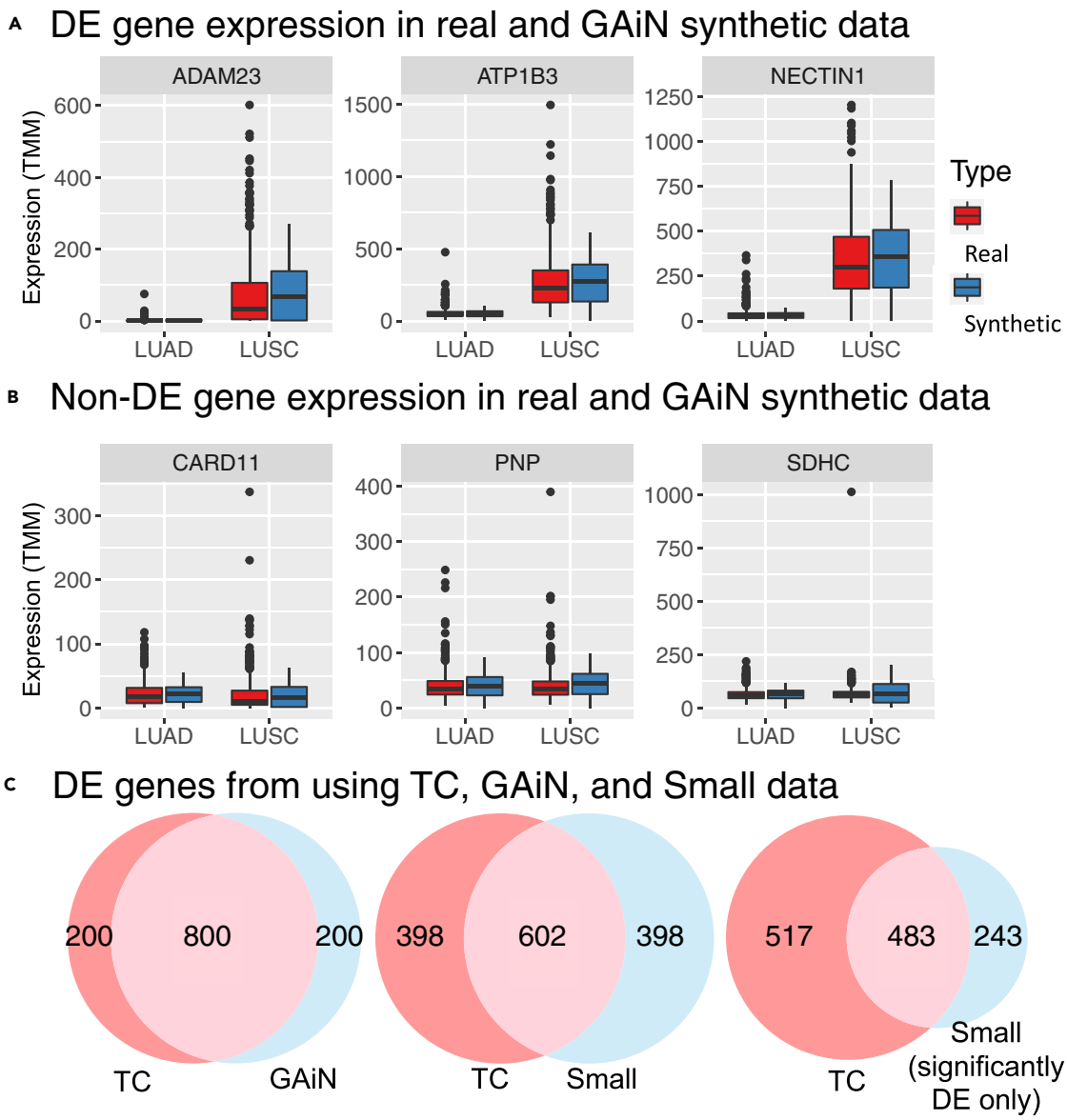


Figure 2. Application of GAiN to TCGA lung cancer phenotypes

We assessed GAiN's performance using TMM-normalized gene expression data from two TCGA lung cancer phenotypes, lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). As in Figure 1D, the combined data from the phenotypes was split into a population cohort (PC) and a test cohort (TC), with a small cohort of 20 samples, 10 from each phenotype, randomly selected from the PC. GAiN was then run with the small cohort samples as input.

(A) Comparison of the TMM-normalized expression of the top 3 genes DE between phenotypes in the TC shows their expression in synthetic samples of the GAiN-generated cohort (synthetic) closely match that of real samples in the TC (real).

(B) Comparison of the 3 least DE genes from the TC shows similar agreement.

(C) Comparison of the TC's true-positive DE genes (red) with DE genes identified by comparison of GAiN augmented cohorts and of the un-augmented small cohort (blue).

numbers, whereas using traditional biostatistical methods can lead to misleading conclusions or false negatives.

Discovery of pathways reflective of known differences between phenotypes

Last, we wished to highlight the ability of GAiN to identify more granular biological pathways when analyzing samples where classical DE and PEA return minimal results. To this end, we separated TCGA-LUSC samples into high (top 25% of the cohort) and low expressers (bottom 25% of the cohort) of the T cell tolerance modulator PDL1 (CD274). Using 5 samples per condition, we repeated our analysis using GAiN and classical DE and PEA. Classical PEA identified a single pathway as significant (Figures 5A and 5D), whereas GAiN augmentation identified approximately 100 enriched pathways, many of which involve immune modulation and immune tolerance, as is the putative

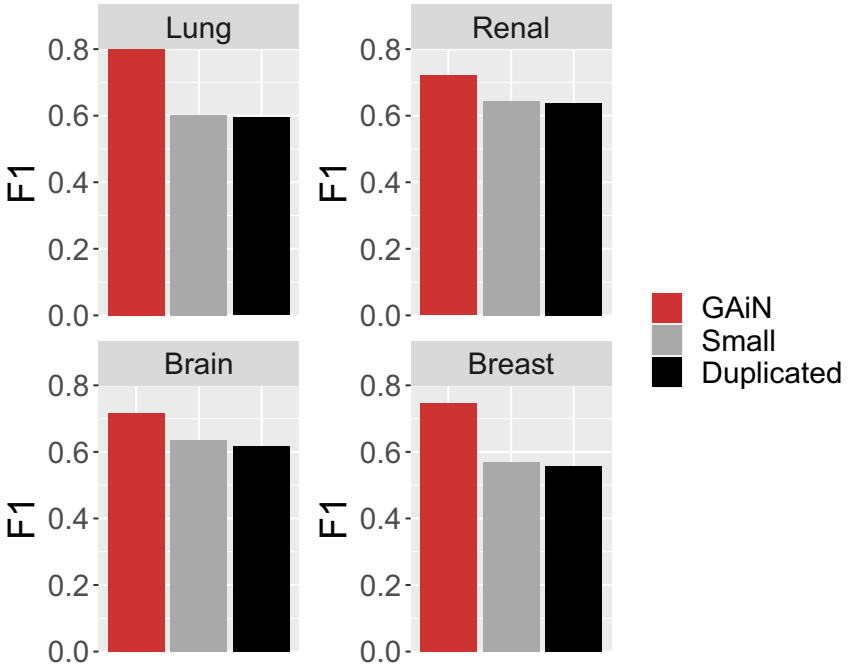


Figure 3. Performance of GAiN in detecting DE genes

The ability of samples generated by GAiN after training on small cohorts with only 10 samples in each phenotype to recapitulate gold-standard DE gene lists obtained from large test cohorts (TCs) (Figure 1D) is quantified by F1 scores. The performance of DE gene analysis on the un-augmented small cohort ("small") and on a cohort created by naive duplication of each sample 50 times ("duplicated") when compared with the TC was also assessed. The purpose of this additional comparison of a "duplicated" version of the small cohort to the TC is to demonstrate that simply reducing the alpha for calling DE genes from the small cohort, equivalent to artificially increasing power by duplicating all samples, provides inferior ability to detect true-positive DEGs from small data compared with GAiN's generative modeling approach. Comparisons were conducted between lung cancer phenotypes (lung adenocarcinoma [LUAD] and lung squamous cell carcinoma [LUSC], "lung"), between renal cancer phenotypes (renal papillary carcinoma [KIRP] and kidney clear cell carcinoma [KIRC], "kidney"), between brain cancers of similar histology but different pathological stage (low-grade glioma [LGG] and glioblastoma [GBM], "brain"), and between triple negative and luminal B subtypes of breast cancer (BRCA, "breast"). In each case, GAiN outperforms the alternatives.

function of PDL1. Other pathways identified by GAiN as significantly enriched between the groups highlight specific cytokine pathways, signaling pathways, and antigen presentation pathways that were not discoverable using a classical workflow on this small sample set (Figures 5B and 5D). These results highlight the utility of GAiN augmentation in downstream analysis of cohorts with limited samples.

DISCUSSION

Reliable analyses of gene expression data using conventional methods are limited when sample numbers are small.[2,9] Low sample numbers hinder the validity and reproducibility of experiments using high-throughput sequencing data. Investigators studying rare disease or rare subtypes of disease often have to wait years to obtain the necessary number of samples for a robust analysis, and in some contexts, this is not an option at all. As a result, investigators studying unique clinical scenarios can lack for study power. Several illustrating examples include patients with HPV^+ head and neck cancer (a good prognostic sign, which often makes a patient a candidate for treatment de-escalation) who have a poor response to chemoradiation therapy (which occurs in approximately 10% of HPV^+ head and neck squamous cell carcinoma [HNSCC] patients).[21] Another example is studying the increasingly iterative and combinatorial approach to locally advanced non-small-cell lung cancer.[22] Current standard of care includes many combinations and sequences of chemotherapy, radiation therapy, surgery, and immunotherapy—in this case and in similar clinical scenarios, finding homogeneously treated cohorts for a robust molecular analysis of the individual effects of each treatment combination is extraordinarily challenging. Additionally, especially in the context of human samples, new treatments (as in phase I clinical trials) are tested which represent a novel clinical context. By definition, novel treatment strategies start with a limited number of samples, and using AI techniques to study emerging technologies allows us to move the field forward more rapidly.

Recent strides in the field of generative modeling, particularly with respect to deep learning GANs, have made it possible to build large augmented synthetic RNA expression datasets for downstream analysis from sparse training data.[23] Herein, we report the design and implementation of an analysis tool, GAiN, that is able to use sparse training data to uncover structural gene expression patterns of phenotypic groups. To our knowledge, this is the first study to use a supervised strategy of separately trained GANs to generate synthetic gene expression profiles for downstream analysis of differential gene expression and pathway enrichment. We demonstrated that GAiN's architecture and hyperparameters were able to reproduce lists of DEGs using less than 2% (i.e., n = 10) of original robust comparisons of cancers of the same organ, cancers of similar origin but different pathological stage, and tumors of similar histology but different subtype. These validation experiments demonstrated that the GAiN workflow is generalizable and its conclusions are not specific to TCGA lung cancer data that the algorithm was trained on, but are also able to predict gene changes in a wide range of scenarios. Although GAN models are usually extremely hard to train, our GAiN framework, tuned on gene expression data, can be easily adopted by researchers even without a background in AI technologies. Our GAiN tool is freely available under the MIT license, and AI researchers can potentially expand its workflow to additional biomedical research applications.

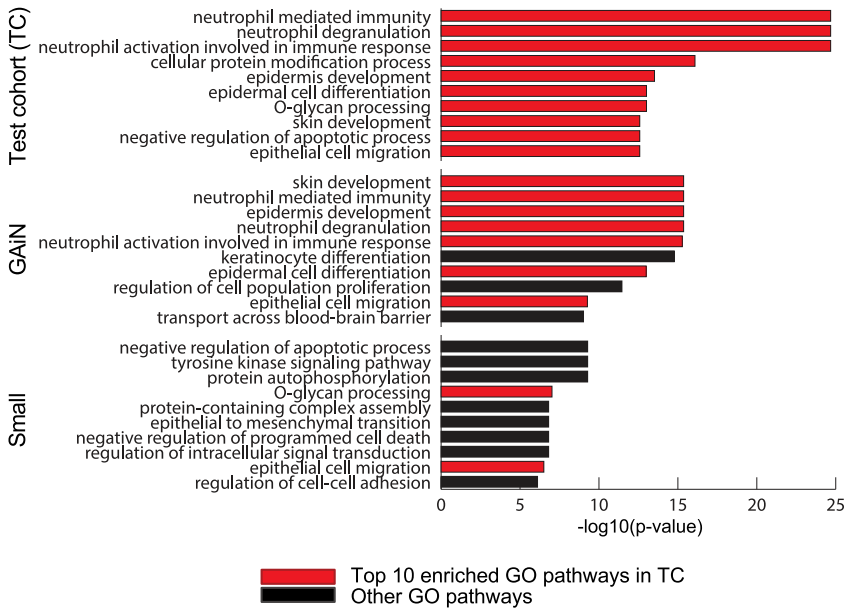
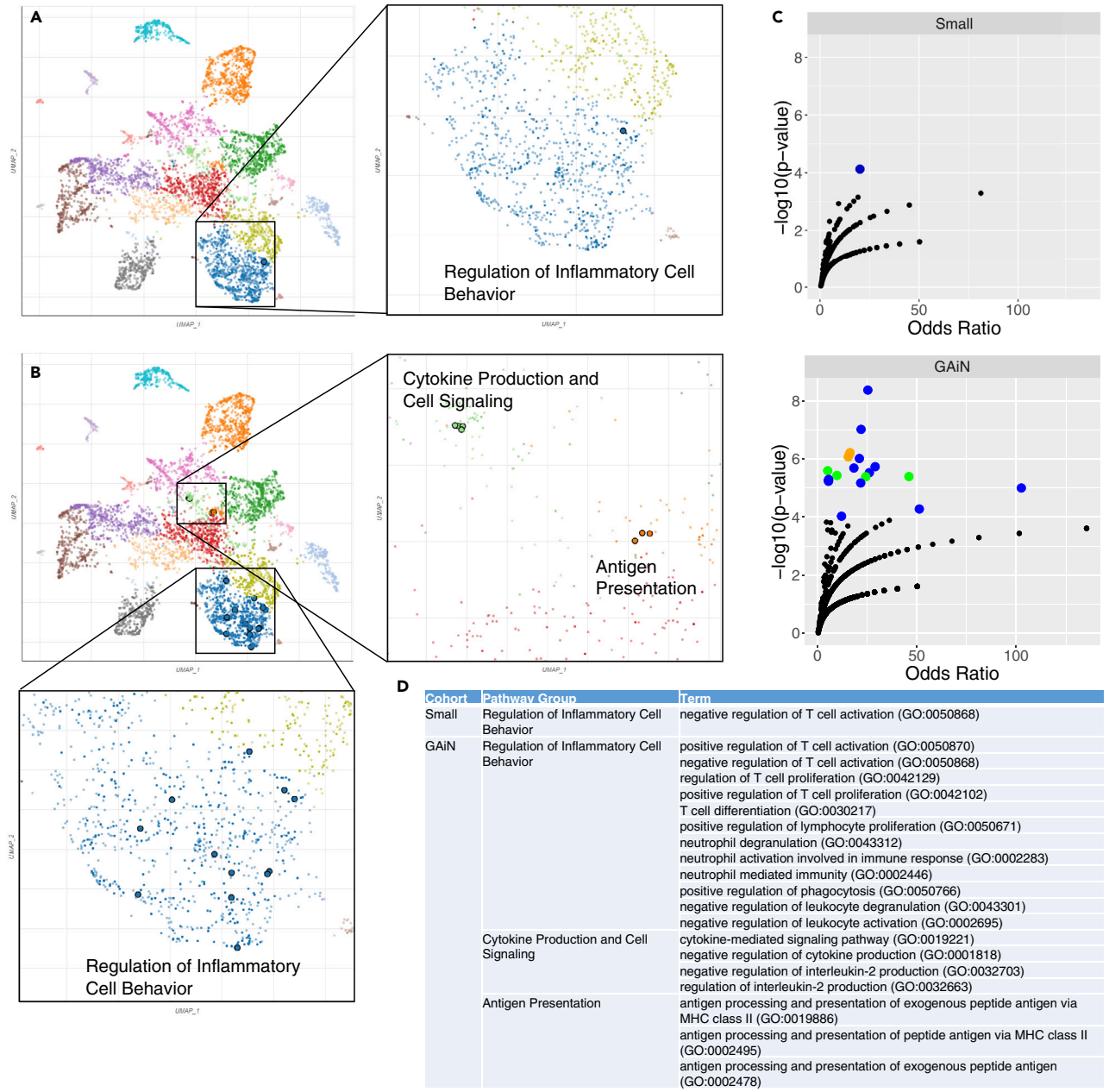


Figure 4. Performance of GAiN in identifying enriched pathways

Top 10 enriched GO terms when performing pathway enrichment analysis (PEA) on the top 1,000 DE genes from the TC (top), GAiN augmented cohort (middle), and small cohort (bottom) when comparing LUAD and LUSC phenotypes. The x axis represents the Benjamini-Hochberg corrected p value from Fisher's exact test as reported by EnrichR for each gene set. Pathways from the top 10 in the gold-standard TC are highlighted in red; 7 of the top 10 enriched GO pathways from the GAiN gene list match the gold-standard TC list, while only 2 of 10 of the PEA results from the non-augmented DE genes match.

To evaluate the ability of GAiN to capture structural gene expression patterns of phenotypic groups, we used TCGA consortium datasets, where large cohorts of hundreds of samples are available for each phenotypic group. However, in contrast to several publicly available sets of large consortium data (e.g., TCGA), the majority of institutional datasets generated in the research community (e.g., GEO, database of Genotypes and Phenotypes [dbGaP]) are usually small, typically with 10–20 samples or fewer in each group (see Figure S1). In addition, because of a commitment to protect participants' information, associated rich clinical data are usually not shared, so the translational aspects of such experiments could be carried out using only a single-institution small sample set. Not surprisingly, conventional biostatistics methods and data analysis approaches (e.g., t test, edgeR, and enrichR) perform poorly when applied directly to such datasets because of the noise contained in small numbers of samples. This can be exemplified by the result that the gene list which passed stringent multiple hypothesis standards in a sparse comparison of LUAD and LUSC data poorly reflected that of the gold standard cohort (Figure 2C) and that enrichR pathway analysis of sparse data discovered none of the top 6 true-positive enriched pathways discovered in the TC (Figure 4). Conversely, using GAiN on the same 10 pairs of lung cancer data, we achieved >93% overall accuracy in discovering DEGs and identified the 5 of the top 6 true-positive enriched pathways in subsequent pathway analysis, all while the individual gene level expression profiles representative of each TC phenotype were largely (although not entirely) preserved (Figures 2A, 2B, S2, S3, and S7). Our benchmarking results demonstrated that GAiN analysis greatly improves on classical analysis of gene expression in a variety of contexts when sample sizes are limited. This finding has broad implications, from empowering research groups with limited resources to more robustly contribute to biological questions using high-throughput informatics, to making subgroup analysis possible in poorly sampled conditions or populations, performing more granular studies on subcohorts (which may be small) of large data, or simply increasing the validity of previous studies which use RNA sequencing (RNA-seq) data with limited sample sizes.

GAiN's ability to robustly analyze small data necessitates inquiry into the reason why the GAN models are effective at predicting true-positive gene expression. Although repeated biological samples are of unquestionable value, technical variability exists in procuring expression measurements.[12,13,24,25] Additionally, computational analysis workflows have differing strengths and biases in quantification of RNA-seq gene counts.[12,26] These measurements additionally do not take into account operator error or marginal variations in sample procurement.[27,28] It follows that in the setting of modest resources, noise inherent in generating an RNA-seq dataset can skew conclusions, especially when observations are sparse. Interestingly, the generator arm of the GAN has the architecture of a DAE. The concept of denoising was conventionally used to delineate signal from noise in image processing, and DAEs have recently shown promise as scalable methods for reducing error in single-cell RNA-seq (scRNA-seq) datasets.[12,23] Although some have postulated the need for additional algorithms to control for technical variation in single cell RNA-seq data where cell types can be sparse, less attention is focused on classical bulk RNA-seq datasets where observations are limited.[13,23,29] Although it was usually assumed in conventional analyses that RNA-seq datasets represent a noiseless representation of the condition they seek to measure, in reality datasets contain a corrupted representation of biology due to errors in the procuring technology. The dense network of the generator arm in GAN architecture relies on the correlational structure of gene expression data to infer "corrected gene expression values." Furthermore, given the architecture of the neural network, DAE architectures are far more scalable and not reliant on linear modeling methods, as are other gene expression imputation techniques.[23] To further increase robustness, our GAiN tool includes a default of 5 GANs, each with a different nodal architecture from the others, and genes are ranked by assembling the outputs from all the GANs (see experimental procedures). Overall, downstream analysis of synthetic RNA-seq data generated using our GAiN tool is therefore not subject to the noise from individual measurements likely experienced in the analysis of a small cohort of RNA-seq



D

Cohort	Pathway Group	Term
Small	Regulation of Inflammatory Cell Behavior	negative regulation of T cell activation (GO:0050868)
GAiN	Regulation of Inflammatory Cell Behavior	positive regulation of T cell activation (GO:0050870)
		negative regulation of T cell activation (GO:0050868)
		regulation of T cell proliferation (GO:0042129)
		positive regulation of T cell proliferation (GO:0042102)
		T cell differentiation (GO:0030217)
		positive regulation of lymphocyte proliferation (GO:0050671)
		neutrophil degranulation (GO:0043312)
		neutrophil activation involved in immune response (GO:0002283)
		neutrophil mediated immunity (GO:0002446)
		positive regulation of phagocytosis (GO:0050766)
		negative regulation of leukocyte degranulation (GO:0043301)
		negative regulation of leukocyte activation (GO:0002695)
	Cytokine Production and Cell Signaling	cytokine-mediated signaling pathway (GO:0019221)
		negative regulation of cytokine production (GO:0001818)
		negative regulation of interleukin-2 production (GO:0032703)
		regulation of interleukin-2 production (GO:0032663)
	Antigen Presentation	antigen processing and presentation of exogenous peptide antigen via MHC class II (GO:0019886)
		antigen processing and presentation of peptide antigen via MHC class II (GO:0002495)
		antigen processing and presentation of exogenous peptide antigen (GO:0002478)

Figure 5. Detection of relevant enriched pathways between high and low PDL1 expressors

A small cohort was constructed by randomly selecting 5 samples from the top 25% of PDL1 expressors and 5 samples from the bottom 25% of PDL1 expressors in the TCGA-LUSC cohort.

(A) Small cohort: uniform manifold approximation and projection (UMAP) of GO pathways with the single pathway identified by EnrichR as enriched with FDR $\leq$ 0.05 between the high and low PDL1 groups highlighted; it falls within the cluster of pathways related to regulation of inflammatory cell behavior (blue).

(B) GAiN cohort: UMAP of GO pathways with the 19 biologically important pathways identified by EnrichR as enriched with FDR $\leq$ 0.05 between the high and low PDL1 groups highlighted; they fall within the clusters related to regulation of inflammatory cell behavior (blue), cytokine production and cell signaling (green), and antigen presentation (orange).

(C) Volcano plots of enriched pathway p value and odds ratio for the small and GAiN cohorts; pathways with FDR $\leq$ 0.05 are highlighted using the same colors as in (B).

(D) Table of the GO pathways enriched with FDR $\leq$ 0.05 in the small and GAiN cohorts. Much more of the biological differences resulting from differential expression of PDL1 become discoverable under GAiN analysis than can be found working with un-augmented small data.

samples with solely biological replicates. To connect GAN output to standard downstream analysis, we converted the standardized values used in the models into trimmed mean of *M* (TMM) values, which is a gene expression normalization method broadly adopted by the community.[30] Additionally, population data, not necessarily of the same phenotype in one's study, can be used to better estimate mean and SD values using a random forest model (see experimental procedures), rather than using the noisier values from the small training cohort itself. If other normalization methods are needed (e.g., relative log expression [RLE], median ratio normalization [MRN]), the code of our GAiN tool can be easily adapted to incorporate them. In addition, we expect that the original standardized values could potentially be used directly in future applications, especially in the area of AI related approaches.

Taken together, we developed an accurate deep learning tool, GAiN, that robustly augments gene expression differences between known phenotypes of limited sample size. We have also demonstrated that phenotypic subgroups have structural gene expression differences which can be uncovered using GAiN analysis. Our publicly available and user-friendly tool can easily be adopted by researchers in the community and applied to their existing datasets or new datasets. Use of GAiN augmented data can be accompanied by classical molecular techniques to confirm patterns one would observe when they collected hundreds or thousands of biological samples. Through our analysis, we demonstrated that our innovative deep learning tool, GAiN, can be applied in a wide array of contexts, and thus has the potential to broadly affect both medical and basic research initiatives.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Jin Zhang (jin.zhang@wustl.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

This paper analyzes existing, publicly available data from the Genomic Data Commons (GDC). The TCGA cohorts are TCGA-LUAD, TCGA-LUSC, TCGA-KIRP, TCGA-KIRC, TCGA-LGG, TCGA-GBM, and TCGA-BRCA.

All original code is available at GitHub (https://github.com/jin-wash-u/GAiN), has been deposited at Zenodo under the document object identifier (DOI) https://doi.org/10.5281/zenodo.10027883, and is publicly available as of the date of publication.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Overview of the GAiN tool

We implemented our data augmentation tool, GAiN, and trained it on small numbers of samples from two separate phenotypes (i.e., LUAD [n = 10] and LUSC [n = 10]) to identify gene expression differences which one would expect to observe only if they collected hundreds or thousands of biological samples using a traditional analysis. To achieve this, GAiN is designed to train two sub-models, a generative model and a discriminative model, for each phenotypic group (Figure 1). Our tool was implemented using the Python programming language, with the sub-models defined using the Keras package sequential application programming interface (API).[31] The generative model includes a DAE architecture that captures the data distribution of the training data. The discriminative model includes a multilayer perceptron architecture that estimates the probability that the sample is drawn from the training data rather than the generative model. Following training of the generators, each generator was used to generate 500 *Z*-normalized synthetic samples for their corresponding phenotype. Scale was restored to gene expression values by estimating the mean and SD of each gene using population data. This PC should ideally match the biology of the phenotypes of interest as closely as possible, but is able to be heterogeneous, containing samples of other phenotypes. A random forest classifier was trained using the small training data and then used to classify the "population data cohort" samples as more like phenotype 1 or phenotype 2. The *Z*-normalized synthetic gene expression values were restored to TMM values normalized values using the gene mean and SD of population samples assigned to the corresponding phenotype.[30] The augmented datasets were then analyzed for DEGs, altered pathways, and other downstream analyses (Figure 1D). The GAiN tool for supervised gene expression analysis is freely available at https://github.com/jin-wash-u/GAiN.

Datasets and preprocessing

TCGA RNA-seq data of solid tumor samples were obtained from the GDC and TMM normalized. To compare cancers from the same organ with different histology, we downloaded transcript abundance estimates for TCGA-LUAD (n = 533), TCGA-LUSC (n = 502), TCGA-KIRP (n = 288), and TCGA-KIRC (n = 538). To compare cancer subtypes, we downloaded gene expression data from TCGA-BRCA cohort, including the most numerous subtype, luminal B (n = 207), and the triple-negative (n = 194) subtype. To compare cancer grades, we downloaded gene expression data from TCGA-GBM (n = 156) and TCGA-LGG (n = 511). Additionally, to show pathway effects of single gene alterations high and low expressors of CD274 (PDL1) in TCGA-LUSC were binned and compared (n = 125). In each study genes were retained for further analysis if they met an average minimum expression and average deviation cutoff of 10.

Training of GAiN sub-models

The two sub-models of GAiN are simultaneously trained via an adversarial process as has been described previously.[32–35] As GAN optimization is notoriously difficult to achieve, we adopted Wasserstein GANs (WGAN) for our strategy, which train the generator model to more closely mirror the training data distribution.[32,36] Briefly, WGANs seek to optimize critic scores which minimize the distance between the distributions of the real and synthetic gene expression data. In order to increase the robustness and stability of the GAiN tool, the final output of GAiN was the average gene rank of 20 separately trained generator-discriminator networks all with different nodal architecture. The latent space of the generator was 128 input nodes with subsequent nodal numbers of hidden layers being $(10 + i \times 100)$, $(10 + i \times 100) \times 2$, $(10 + i \times 100) \times 3$, and gene vector length, for all networks i from 1 to 20. Similarly, the discriminator nodal densely connected hidden layer architecture for the ith network was $(10 + i \times 100) \times 3$, $(10 + i \times 100) \times 2$, and $(10 + i \times 100)$. For every layer of the generator and discriminator network, a Leaky ReLU activation function was used in an effort to prevent dead gradients. Dropout was used at multiple levels of the discriminator and the generator to prevent overfitting (for full model architecture the Python code has been made publicly available as indicated above). The generator and the discriminator were trained in an alternating fashion as described in the original GAN manuscript.[32] Performance metrics for training GAiN are presented in Table S3.

Network compilation and fit

Additional WGAN hyperparameters, including the optimization algorithm, learning rate, and critic clipping, were set as recommended in previous WGAN studies.[33,34] The RMSProp optimization algorithm was used with a learning rate of 5.0×10^{-6}. The Wasserstein loss function served as the loss function to be minimized, critic clipping was set at 0.01. Each network was trained for 500 epochs (empirically determined to result in convergence for most nodal architectures).

Restoration of gene expression values using population data

In order to restore scale to *Z*-normalized synthetic samples, gene mean and SD were estimated from population data. Population data were separated from the test dataset used as the gold-standard comparison by a train-test

split of 50%, stratified by phenotype, prior to running GAiN. Sparse training data (n = 10) for each phenotype X and Y were used to train a random forest classifier with 1,000 estimators. Random forest classification was implemented using the scikit-learn Python package. Samples in the PC were classified as phenotype X or Y. *Z*-normalized synthetic gene expression values were restored using the gene mean and SD, estimated from the corresponding population samples, assigned to the phenotype by the random forest classifier.

DE analysis

When using a cohort of large sample size (e.g., n = 500–1,000), a large fraction of the total genes tested were deemed significant by multiple hypothesis testing. For useful discrimination, we used a rank-based method, and tested the top 1,000 DEGs from the GAiN augmented synthetic cohort, true positive TC, and non-augmented small cohort, as determined by the edgeR R package.[16] To illustrate performance when different DEG thresholds are used for comparison, ROC curves were plotted of the true-positive rate (TPR) and false-positive rate (FPR) of GAiN cohort genes compared with TC genes as the number of genes compared varies between 1 and the total number of genes in the dataset.

PEA and comparison

Genes among the top 1,000 DE list in each analysis were passed to the enrichR package.[17] The reference database used in enrichment analysis was the GO database.[18] The top 10 pathways were selected and presented from each enrichment analysis.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2023.100910.

ACKNOWLEDGMENTS

This research was supported by National Cancer Institute grants R01CA276955, R21CA264343, K22CA237839, and R37CA279596 (J.Z.).

AUTHOR CONTRIBUTIONS

Conceptualization, M.R.W. and J.Z.; methodology, M.R.W. and M.I.; investigation, M.R.W., M.I., J.S., R.M.K., C.R., S.J.S., O.L.G., and J.Z.; visualization, M.R.W., M.I., K.J., and J.Z.; supervision, J.K.S. and J.Z.; writing – original draft, M.R.W., M.I., and J.Z.; writing – review & editing, M.R.W., O.L.G., and J.Z.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 20, 2023
Revised: October 23, 2023
Accepted: December 7, 2023
Published: January 8, 2024

REFERENCES

1. Decherchi, S., Pedrini, E., Mordenti, M., Cavalli, A., and Sangiorgi, L. (2021). Opportunities and Challenges for Machine Learning in Rare Diseases. Front. Med. *8*, 747612.
2. Ioannidis, J.P.A. (2005). Why Most Published Research Findings Are False. PLoS Med. *2*, e124.
3. Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., and Munafò, M.R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. *14*, 365–376.
4. Khatri, P., Sirota, M., and Butte, A.J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Comput. Biol. *8*, e1002375.
5. Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., et al. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat. Protoc. *14*, 482–517.
6. Zuin, J., Roth, G., Zhan, Y., Cramard, J., Redolfi, J., Piskadlo, E., Mach, P., Kryzhanovska, M., Tihanyi, G., Kohler, H., et al. (2022). Nonlinear control of transcription through enhancer–promoter interactions. Nature *604*, 571–577.
7. Regondi, C., Fratelli, M., Damia, G., Guffanti, F., Ganzinelli, M., Matteucci, M., and Masseroli, M. (2021). Predictive modeling of gene expression regulation. BMC Bioinf. *22*, 571.
8. Schwabe, A., Rybakova, K.N., and Bruggeman, F.J. (2012). Transcription stochasticity of complex gene regulation models. Biophys. J. *103*, 1152–1161.
9. Stretch, C., Khan, S., Asgarian, N., Eisner, R., Vaisipour, S., Damaraju, S., Graham, K., Bathe, O.F., Steed, H., Greiner, R., and Baracos, V.E. (2013). Effects of sample size on differential gene expression, rank order and prediction accuracy of a gene signature. PLoS One *8*, e65380.
10. Baccarella, A., Williams, C.R., Parrish, J.Z., and Kim, C.C. (2018). Empirical assessment of the impact of sample number and read depth on RNA-Seq analysis workflow performance. BMC Bioinf. *19*, 423.
11. Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. *1*, 207–210.
12. Eraslan, G., Simon, L.M., Mircea, M., Mueller, N.S., and Theis, F.J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. Nat. Commun. *10*, 390–414.
13. Le, H.-S., Schulz, M.H., McCauley, B.M., Hinman, V.F., and Bar-Joseph, Z. (2013). Probabilistic error correction for RNA sequencing. Nucleic Acids Res. *41*, e109.
14. Inkman M. (2023). jin-wash-u/GAiN: GAiN. Version 1.0. Zenodo; . https://doi.org/10.5281/zenodo.10027883.
15. Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113–1120.
16. McCarthy, D.J., Chen, Y., and Smyth, G.K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. *40*, 4288–4297.
17. Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. *44*, W90–W97.
18. Gene Ontology Consortium (2021). The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. *49*, D325-d334.
19. Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. *28*, 27–30.
20. Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C., et al. (2022). The Reactome Pathway Knowledgebase 2022. Nucleic Acids Res. *50*, D687–D692.
21. Yoo, S.H., Ock, C.Y., Keam, B., Park, S.J., Kim, T.M., Kim, J.H., Jeon, Y.K., Chung, E.J., Kwon, S.K., Hah, J.H., et al. (2019). Poor prognostic factors in human papillomavirus-positive head and neck cancer: who might not be candidates for de-escalation treatment? Korean J. Intern. Med. (Engl. Ed.) *34*, 1313–1323.
22. Antonia, S.J., Villegas, A., Daniel, D., Vicente, D., Murakami, S., Hui, R., Yokoi, T., Chiappori, A., Lee, K.H., de Wit, M., et al. (2017). Durvalumab after Chemoradiotherapy in Stage III Non–Small-Cell Lung Cancer. N. Engl. J. Med. *377*, 1919–1929.
23. Wang, J., Agarwal, D., Huang, M., Hu, G., Zhou, Z., Ye, C., and Zhang, N.R. (2019). Data denoising with transfer learning in single-cell transcriptomics. Nat. Methods *16*, 875–878.

24. Lahens, N.F., Kavakli, I.H., Zhang, R., Hayer, K., Black, M.B., Dueck, H., Pizarro, A., Kim, J., Irizarry, R., Thomas, R.S., et al. (2014). IVT-seq reveals extreme bias in RNA sequencing. Genome Biol. *15*, R86.

25. Mai, Z., Xiao, C., Jin, J., and Zhang, G. (2017). Low-cost, low-bias and low-input RNA-seq with high experimental verifiability based on semiconductor sequencing. Sci. Rep. *7*, 1053–1063.

26. Chen, L.Y., Wei, K.-C., Huang, A.C.-Y., Wang, K., Huang, C.-Y., Yi, D., Tang, C.Y., Galas, D.J., and Hood, L.E. (2012). RNASEQR—a streamlined and accurate RNA-seq sequence analysis program. Nucleic Acids Res. *40*, e42.

27. Tong, L., Yang, C., Wu, P.-Y., and Wang, M.D. (2016). Evaluating the impact of sequencing error correction for RNA-seq data with ERCC RNA spike-in controls. IEEE. EMBS. Int. Conf. Biomed. Health Inform. *2016*, 74–77.

28. Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., and Heisler, M.G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods *10*, 1093–1095.

29. Shao, L., Yan, R., Li, X., and Liu, Y. (2014). From heuristic optimization to dictionary learning: A review and comprehensive comparison of image denoising algorithms. IEEE Trans. Cybern. *44*, 1001–1013.

30. Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25.

31. Chollet, F.; Keras Team (2015). Keras. https://keras.io.

32. Arjovsky, M., and Bottou, L. (2017). Towards principled methods for training generative adversarial networks. Preprint at arXiv.

33. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A.C. (2017). Improved training of wasserstein gans. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17) (ACM), pp. 5767–5777.

34. Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein GAN. Preprint at arXiv.

35. Islam, J., and Zhang, Y. (2020). GAN-based synthetic brain PET image generation. Brain Inform. *7*, 3.

36. Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. Preprint at arXiv.