Protein structure–based gene expression signatures

Rayees Rahman^{a,1} , Nicole Zatorski^a , Jens Hansen^{a,b} , Yuguang Xiong^{a,b} , J. G. Coen van Hasselt^a , Eric A. Sobie^{a,b} , Marc R. Birtwistle^{a,c} , Evren U. Azeloglu^{a,d} , Ravi Iyengar^{a,b,1} , and Avner Schlessinger^{a,1}

^aDepartment of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029; ^bInstitute for Systems Biomedicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029; ^cDepartment of Chemical and Biomolecular Engineering, Clemson University, Clemson, SC 29634; and ^dDivision of Nephrology, Icahn School of Medicine at Mount Sinai, New York, NY 10029

Edited by Roland L. Dunbrack, Jr., Fox Chase Cancer Center, Philadelphia, PA, and accepted by Editorial Board Member Michael Levitt March 27, 2021 (received for review July 16, 2020)

Gene expression signatures (GES) connect phenotypes to differential messenger RNA (mRNA) expression of genes, providing a powerful approach to define cellular identity, function, and the effects of perturbations. The use of GES has suffered from vague assessment criteria and limited reproducibility. Because the structure of proteins defines the functional capability of genes, we hypothesized that enrichment of structural features could be a generalizable representation of gene sets. We derive structural gene expression signatures (sGES) using features from multiple levels of protein structure (e.g., domain and fold) encoded by the mRNAs in GES. Comprehensive analyses of data from the Genotype-Tissue Expression Project (GTEx), the all RNA-seq and ChIP-seq sample and signature search (ARCHS4) database, and mRNA expression of drug effects on cardiomyocytes show that sGES are useful for characterizing biological phenomena. sGES enable phenotypic characterization across experimental platforms, facilitates interoperability of expression datasets, and describe drug action on cells.

gene expression signatures | structural bioinformatics | reproducibility

Gene expression signatures (GES) are generally defined as a ranked list of genes whose differential expression is associated with a particular biological phenomenon (1–3). GES are typically obtained by measuring the transcriptional level of genes through RNA sequencing or previously by array-based experiments. Often, gene sets in GES are heuristically determined, for example, by taking the top 100 to 250 highly expressed genes or by using particular *p*-value cutoffs (2).

Thousands of GES have been used to describe a wide variety of biological phenomenon (1–3). For example, GES were used to characterize subcellular and whole-cell functions (4, 5), pathological states (6, 7), and cellular response to various perturbagens such as drugs (8). In addition, multiple resources of differential gene expression data have been made available to the scientific community (9, 10), including the Genotype-Tissue Expression Project (GTEx) (11) and the all RNA-seq and ChIP-seq sample and signature search (ARCHS4) database, a collection of tissue expression data mined from the Gene Expression Omnibus (12). These databases are commonly used to analyze tissue-specific transcriptomics data for different purposes (13, 14).

Determining the reproducibility of a signature for a phenotype of interest remains a challenge, often requiring meta-analyses of existing GES to validate a signature for a phenotype (3, 15). This process includes analyses of thousands of independent samples to generate a robust signature for a single phenotype (16). Because of differences in technology, normalization protocols, and practices across laboratories, there is variability in identifying robust GES for a given phenotype, which has hindered GES utility in the clinic (15, 17). GES variability has led to the cancellation of clinical trials that linked endpoints to specific GES and can produce inconsistent results in the classification of patients for distinct subtypes of a cancer (15, 18). For example, Patil et al. observed that depending on the control used to compute differential gene expression, GES can produce differing breast cancer subtype predictions for a single patient (15).

Recent studies have demonstrated the requirement of large, multiscale analyses to determine the reproducibility of GES data across studies (15, 18–22). Ein-Dor et al. introduced measuring the overlap of GES across two datasets profiling the same phenotype as a measure of robustness of the GES (16). They estimate that for a pair of breast cancer GES of size 70, to achieve a minimal overlap of 50% between the GES would require 2,300 gene expression samples from patients. Dhawan et al. developed sigQC, a protocol for quantitatively evaluating previously obtained GES across multiple gene expression databases (23). However, determining robustness and reproducibility of GES for newly obtained transcriptomics data remains challenging.

One way to improve the robustness of GES is to integrate multiple types of biological information (1, 24). Set2Gaussian embeds GES with a protein–protein interaction network into a low-dimensional feature space that improves gene set member identification (25). Because genes encode proteins with three-dimensional structures, we hypothesize that identifying enriched features derived from protein structure define a generalizable representation of a gene set and may provide mechanistic and evolutionary insights not captured by expression patterns alone. Particularly, one common structural characteristic of proteins is the overall structural family or “fold” of the protein and/or its individual domains. The structural fold or folds of a protein capture evolutionary relationships between proteins (and cognate) genes as well as protein modularity; therefore, they have a direct association with gene evolution and physiological function (26–28). Incorporating structural information has enhanced the

Significance

Gene expression signatures (GES) are used to define phenotypes, such as cell and tissue state, or the effects of perturbagens. GES have suffered from vague assessment criteria and limited reproducibility. Here, we show that combining protein structure features with GES using deep learning approaches significantly enhances the reliability of the transcriptomic signatures across datasets representing various phenotypes, including relationships among tissues and drugs.

Author contributions: R.R., R.I., and A.S. conceptualized the study; R.R. curated the data, performed the formal analysis, developed the methodology for structural signatures, and trained and validated the machine learning models used in the analysis; N.Z. evaluated the sGES method over several parameters; Y.X. performed the transcriptomics analysis of the Promocell data; E.A.S., M.R.B., and E.U.A. provided critical analyses and insight into study design; R.R. and A.S. wrote the analysis; R.R., J.H., J.G.C.v.H., E.A.S., M.R.B., E.U.A., R.I., and A.S. edited and reviewed the analysis; and R.I. and A.S. provided funding for the project.

Competing interest statement: R.R. and A.S. are co-founders of Aichemy Inc.

This article is a PNAS Direct Submission. R.L.D. is a guest editor invited by the Editorial Board.

Published under the PNAS license.

¹To whom correspondence may be addressed. Email: rayees.rahman@icahn.mssm.edu, ravi.iyengar@mssm.edu, or avner.schlessinger@mssm.edu.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2014866118/-DCSupplemental>.

Published May 3, 2021.

prediction of protein–protein interaction networks and disease pathways (29, 30).

In this study, we use structural information present in gene expression profiles to improve reproducibility of GES. We first identify the structural features (e.g., folds and domains) from ranked gene lists to yield structural GES (sGES) and evaluate whether representing sGES as enrichments of structural features retains functional and biological relationships between tissue types. Next, we define measures to evaluate GES and sGES in terms of consistency and predictivity within these databases and measure signature robustness across GTEx and ARCHS4. We then integrate sGES with GES through a machine learning architecture, called an autoencoder, to identify outlier samples, facilitating interoperability between experiments that use differing

transcriptomic methodologies. We demonstrate the generalizability of sGES by characterizing distinct phenotypes, such as cellular response to drugs. These analyses indicate that sGES enables a multidimensional characterization of cellular phenotypes.

Results

Protein Structure Enrichment Captures Relevant Biological Information from Gene Expression Data. Protein structures encoded by the expressed genes in a GES determine a signature’s phenotype. Therefore, we hypothesize that using features derived from protein structures in GES, which, implicitly, have direct association with function, can improve reproducibility across experimental platforms. An sGES for each gene set is determined by identifying structural features from the encoded protein of each gene

Fig. 1. Protein structure enrichment clusters tissue-specific gene expression. (A) Structural Classification of Proteins—extended (SCoPE) hierarchy of protein structural features, with examples. (B) Workflow to generate sGES. (C) The top 250 highest-expressed genes from GTEx (in terms of transcripts per million) were obtained. Tissue samples were clustered based on the presence or absence of the GES using t-SNE. sGES were then derived from the GES, and tissue samples were clustered by using t-SNE based on the presence or absence of structural features at the domain and fold levels. Each sample is colored by tissue type.

(*Materials and Methods* and Fig. 1 *A* and *B*) in the signature. We define a structural feature as the predicted structure (i.e., family, superfamily, fold, and domain) of each gene product based on their sequence similarity to known protein structures.

Specifically, for each gene, we first identify structures of homolog structures using HHpred (31). To assign a “structural feature” we use the structural hierarchy from the Structural Classification of Proteins—extended (SCOPe) (32) and InterProScan (33) databases, in which protein domains (10,637 domains; e.g., serine–threonine/tyrosine–protein kinase and catalytic domain) are categorized into families (4,919 families, e.g., protein kinases and catalytic subunit), which are categorized into superfamilies (2,026 superfamilies; protein kinase–like) and further grouped into distinct folds (1,232; protein kinase–like) (*Materials and Methods* and Fig. 1*A*). For a given gene set, each structural feature is evaluated for enrichment in the gene set, at each structural level (e.g., fold and domain). The observed structures and their counts in the GES are compared to the counts of the structural feature in the human proteome (Fig. 1*B*). Importantly, genes without any assigned structure are labeled as “no-fold” or “no-domain” and are also tested for enrichment compared to the human proteome. sGES are defined as the complete set of structural features derived from a ranked list of genes at each structural level (domain to fold levels).

Next, we determined the minimal number of genes that are needed to produce a stable and consistent sGES. We use the Jaccard coefficient (J_C), which measures the overlap of members between two distinct sets, similarly to the F-statistic introduced by Ein-Dor et al. (16). Our reasoning is that as GES size increases, there will be rapid changes in the members of the extracted sGES, and these changes will be reflected by fluctuations of pairwise J_C distributions. We visually observed that the overlap between pairs of sGES rapidly fluctuate between GES sizes 0 and 150 and begin to stabilize at a GES of size 250 (*SI Appendix*, Fig. S1).

To determine whether protein structure enrichment captures biological information observed in GES, we clustered GTEx tissue samples based on the 250 highest expressed genes, using t-distributed stochastic network embedding (t-SNE) (34) (*SI Appendix*, Table S1 and Fig. S2 and Fig. 1*C*). We observed that sGES are capable of clustering tissue types at both the lowest structural level (domains) and the highest structural level (folds). Notably, the clusters at all structural levels capture functional and spatial relationships among tissues (*SI Appendix*, Fig. S2).

For example, ovary tissue sGES cluster near uterine signatures. Both tissues’ GES are enriched with protein domains related to sex hormone production such as the follistatin/osteonectin epidermal growth factor (EGF) domain, Kazal domain, secreted protein acidic and rich in cysteine (SPARC)/testican, and fibrillar collagen domain (*SI Appendix*, Fig. S3). Both tissues’ sGES also retain tissue-specific domain differences, such as ovarian tissue having sGES-containing glutathione transferase domains, which is a biomarker of oocyte viability and quality (26), while uterine tissue is enriching for proteins containing structural domains, such as the tubulin/FtsZ domain (27).

This result is interesting because conservation of high-level protein structure (i.e., fold) is not necessarily always predictive of protein function (28), yet using a representation of folds, domains, and families, independently, can constitute an sGES that identifies tissue types.

Quantitative Measures to Evaluate GES and sGES Reproducibility. To assess reproducibility of GES of a given phenotype, we defined relevant, quantitative properties to describe the quality of a dataset. We posit that a reproducible GES would have three major properties: 1) consistency across independent samples and replications, 2) high predictive capacity for the phenotype using

standard performance measures (35), and 3) robustness across different measurement platforms.

To evaluate GES consistency, we reuse the J_C , which measures the overlap between two distinct gene sets, characterizing the same phenotype in independent samples (*Materials and Methods* and Fig. 2*A*). A high J_C demonstrates that the GES is consistent across experimental samples. Furthermore, a low variance for a distribution of J_C values across several hundred independent samples may indicate signatures with high reproducibility.

To evaluate the predictive performance of a signature to a phenotype, we use a standardized machine-learning algorithm (a random forest) across all signatures to assess the baseline predictive performance of a given GES in terms of area under the receiver operating characteristic (ROC) curve (AUC) without any significant parameter optimizations or feature selection (Fig. 2*B*).

To measure GES robustness, we computed both J_C and AUC values of signatures measuring the same phenotype across two independent databases. Particularly, we use GTEx and ARCHS4 as two nonoverlapping databases measuring the transcriptional profile of tissues. We first derive GES and sGES based on the 250 highly expressed genes for each tissue type in GTEx. We then measure the consistency and the predictive performance of these signatures within the GTEx dataset. Finally, we measure the robustness of GES and sGES GTEx signatures against the ARCHS4 database (Fig. 2 *B* and *C*).

Consistency and Predictivity of GTEx Tissue Signatures for GES and sGES. We first aimed to determine the average J_C of an inconsistent GES as a negative control. We took equivalently sized sets of highly expressed genes from GTEx (the top 50, 250, and 1,000 highest expressed genes, respectively) and computed the J_C between two randomly selected tissue types. We observed that the average J_C of a pair of inconsistent GESs is 0.33 (Fig. 3*A*).

Next, we tested the consistency of GES and sGES derived from the 250 highest expressed genes for each tissue type in GTEx. Specifically, we computed the J_C values for each pair of samples from the same tissue type, using both GES and sGES, and plotted the distributions of J_C values. Interestingly, consistency of sGES, as measured by J_C , at all structural levels, increases as higher-order sGES are used (Fig. 3*B* and *SI Appendix*, Fig. S4). For a gene set size of 250, sGES significantly increases the mean within-tissue J_C compared to GES across all tissue types (Fig. 3*B* and *SI Appendix*, Table S2). For example, at the fold level, the mean within-tissue J_C reaches a value of 0.75, while the mean across-tissue J_C reaches a value of 0.54 (Fig. 3*B*). The increase in J_C at higher structural levels represents increasing consistency of sGES (Fig. 3*A*).

One explanation for this improvement in consistency is the diminishing number of possible structures as higher-order structural features are used (*SI Appendix*, Fig. S5). However, we observe that while the mean J_C generally increases using sGES, disparate tissue types have significantly lower J_C values at each structural level compared to J_C within the same tissue (Fig. 3*C*). Importantly, the average consistency of dissimilar (or across-tissue) samples using sGES is similar to that of GES (for domain and family levels). This result asserts that while the higher average J_C seen in sGES can be an artifact of decreasing feature space sizes, sGES does retain tissue-specific features that can uniquely identify tissue types (Fig. 1*C*).

After demonstrating that sGES can be used to improve the consistency of gene signatures within GTEx, we evaluated whether using sGES can be used to improve the consistency of GES generated from methods that identify robust GES across several samples. We generated sGES with tissue-specific GES derived from the Molecular Signatures Database (mSigDB) (10), the Human Protein Atlas (36), and TISSUES 2.0 databases (37) (Fig. 3*D*). These approaches use statistical methods and scores to collate several gene expression experiments into a single GES

Fig. 2. Metrics to evaluate GES consistency, predictivity, and robustness. (A) Evaluation metrics for GES consistency, predictivity, and robustness. (B) Approach of measuring consistency, robustness, and outlier detection. (C) Workflow for evaluating the reproducibility of GES, structural signatures, and integrated signatures from GTEx and ARCHS4.

representing a given tissue type. We evaluated the consistency of the GES from these methods to selected GTEx tissue types using pairwise J_c distributions. The average J_c values of GES between each method and GTEx is very low (between 0.01 and 0.15); however, sGES dramatically improves the consistency of these approaches to GTEx (0.25 to 0.50), demonstrating its capability to improve existing approaches to identify robust GES as well as its overall generalizability.

An additional test of GES reproducibility is the ability of a GES, for a phenotype obtained in one sample, to accurately predict the phenotype for a GES derived in an independent sample measuring the same phenotype (i.e., predictivity; Fig. 2A). To evaluate the baseline predictivity of both GES and sGES across different tissues and signature types, we trained a random forest to identify tissues from either GES of size of 250 or sGES from GTEx expression data. The parameters for the random forest were standardized, and neither feature selection nor parameter optimization was performed on any model (*Materials and Methods*, Fig. 4, and *SI Appendix*, Fig. S6).

Both GES and sGES at any structural level (e.g., fold and superfamily) have high predictivity for any given tissue type within the GTEx dataset. For example, the best AUC values for each tissue range from 0.891 (ectocervix) to 1 (lung) (*SI Appendix*, Fig. S6). The tissue with the largest variance in J_c distribution (i.e., ectocervix) has the lowest predictive performance, indicating a relationship between the two evaluation measures. There are small differences in the predictivity among gene set sizes of 50, 250, and 1,000 across all tissue types within GTEx (*SI Appendix*, Fig. S7).

Robustness of GES and sGES across Databases. We then used an independent validation set from ARCHS4 (12) to evaluate the robustness of tissue GES and sGES from GTEx (Fig. 5 and *SI Appendix*, Figs. S8–S11). In brief, ARCHS4 is a collection of gene expression data derived from the Gene Expression Omnibus (38), which collates gene expression data generated from a wide variety of sequencing technologies and platforms. Specifically, we evaluated GTEx signatures for consistency (Fig. 5A and B and *SI Appendix*, Figs. S8 and S9) and predictivity against ARCHS4 (Fig. 5C and D and *SI Appendix*, Figs. S10 and S11).

We derived the 250 highest expressed genes for each tissue type from the ARCHS4 database. We observed that internal ARCHS4 GES consistency is much more variable across tissue types than that of GTEx GES (Fig. 5A, purple, and *SI Appendix*, Fig. S8), likely because of the heterogeneity of the samples in ARCHS4. Samples in ARCHS4 can be obtained from both pathological and healthy tissues or may characterize distinct subtypes of tissues or have artifacts due to differing sequencing methodologies.

To directly measure the robustness of the tissue signatures, we computed pairwise J_c between ARCHS4 and GTEx GES for each tissue type (Fig. 4A, blue). We observe low J_c values (below our consistency threshold of 0.33; Fig. 3A) across most tissue types between ARCHS4 and GTEx GES (Fig. 5A, blue, *SI Appendix*, Fig. S8). Interestingly, for all tissues, sGES at any structural level increases the average pairwise J_c between GTEx and ARCHS4 signatures and thus improves the consistency between the two datasets compared to GES alone (Fig. 5B and *SI Appendix*, Fig. S9). However, for many tissues, the consistency overlap was still close to our J_c baseline value of 0.33.

Fig. 3. Signature consistency improves using protein structure. (A) Pairwise GES J_C distributions across randomly selected, distinct tissues types. The average pairwise, inconsistent GES J_C was determined to be 0.33 across gene set sizes of 50, 250, and 1,000 genes in a GES. (B) Distributions of J_C values within tissue types. For each pairs of samples, in each tissue type (as cataloged by GTEx), a J_C was computed for the top 250 highly expressed genes (by transcripts per million [TPM]) and their derivative sGES at each structural level. All distributions are statistically significant from each other using pairwise t tests, with false discovery rate (FDR) correction (*SI Appendix, Table S2*). The red line indicates a J_C of 0.33, *** $P < 0.001$ from a two-sided t test between each comparison. (C) Distributions of structures are randomly assigned to each gene. "Across Tissues" are J_C distributions between unlike tissue types. "Within Tissues" are the J_C distributions between the same tissue type. Red line indicates a J_C of 0.33, *** $P < 0.001$ from a two-sided t test between each comparison. (D) Evaluation of sGES on mSigDB, The Human Protein Atlas, and TISSUES 2.0 methods of generating robust, tissue-specific GES. Pairwise J_C were generated between tissue-specific signatures from mSigDB and the Human Protein Atlas to the highest expressed genes from GTEx database.

Fig. 4. Predictivity of GES and sGES within the GTEx database. A random forest was trained using GES (of size 250) and sGES at different structural levels (Domain, Family, Superfamily, and Fold) for GTEx tissue expression data. ROC curves are displayed for each structural level.

Next, we aimed to evaluate the predictivity of GES and sGES between orthogonal databases. Surprisingly, there is high predictivity of tissues from ARCHS4 using standardized models trained either on GES or sGES from GTEx (Fig. 5C and D and *SI Appendix*, Figs. S10 and S11). For example, AUC values for each tissue range from 0.70 (pancreas) to 0.999 (vagina) (Fig. 5C and *SI Appendix*, Fig. S10). Importantly, decreasing GES size to 50 genes for many tissue types has significant effects on the performance of the classifier (Fig. 5C and *SI Appendix*, Fig. S9). Several tissues such as pancreas and heart exhibit better performance using a small GES size. This indicates that much of the predictive performance of GES and sGES may be due to a select set of genes and structures rather than the signature as a whole (Fig. 5C and *SI Appendix*, Figs. S10 and S11).

Using both assessment measures, certain tissues such as pancreas, lung, and esophagus are not robust across GTEx and ARCHS4 because of relatively low AUC and J_C values. The only potentially robust signature observed is muscle tissue, where high internal consistency within ARCHS4 and GTEx GES led to a relatively higher overlap J_C distribution across the two datasets (Fig. 5). We hypothesize that identifying and removing pathological and other atypical samples (“outliers”) present in different datasets such as ARCHS4 will improve reproducibility (Fig. 5A).

Integration of sGES and GES Enables Better Outlier Detection. To identify potential outliers in GTEx samples, we used an artificial neural network architecture, termed autoencoder. Autoencoders encode high dimensional data to a lower-dimensional feature space that can regenerate the input of the network (39). The performance of an autoencoder is measured by the reconstruction error between the original inputs and the reconstructed

output. Samples with high reconstruction error are often samples that are considered anomalies or outliers compared to the samples used to train the model.

We trained a stacked denoising autoencoder on 80% of GTEx GES. The remaining 20% of the GTEx GES was used to determine the baseline level of reconstruction error of the autoencoder (Fig. 6A, green, and *SI Appendix*, Fig. S12). We defined samples with reconstruction errors greater than two SDs of the average reconstruction error (0.00725) as outlier samples within GTEx. Notably, based on this definition, very few GTEx samples can be considered outliers.

The autoencoder, when trained with data from GTEx categorized the majority of ARCHS4 samples, within the same tissue type, as “outliers” (Fig. 6A, purple, and *SI Appendix*, Fig. S12). This result corroborates the low consistency between pancreas tissue GES, whose signature was demonstrated to be not robust across ARCHS4 and GTEx (Fig. 5); however, it also categorizes most of ARCHS4 muscle tissue as incompatible with GTEx, despite observing some level of robustness across these databases.

We then evaluated if sGES could improve the identification of outliers and the compatibility of ARCHS4 and GTEx. While we expected outlier detection to be less sensitive by ascending the structural hierarchy, as observed before (Figs. 3–5), surprisingly, distinct levels of structure have differing sensitivity to outliers (Fig. 6B and *SI Appendix*, Fig. S13). For muscle tissue, the family and superfamily levels of sGES identified less outliers than those identified by domain-, fold-, or gene-level signatures. This indicates that distinct levels of the structure hierarchy characterize unique aspects of biological information present in GES.

We hypothesized that integrating GES and sGES would allow us to obtain a consensus classification of outlier versus nonoutlier

Fig. 5. Robustness of GTEx GES using the ARCHS4 database. (A) Distributions of J_c values for a gene signature size of 250 for tissues within the ARCHS4 (purple), the GTEx (green), and across the ARCHS4 and GTEx (blue) databases. Red line indicates a $J_c = 0.33$. (B) Overlap of GTEx sGES with ARCHS4 signatures, across all structure levels. Red line indicates a $J_c = 0.33$. (C) Predictive performance of a random forest model on GTEx gene sets of size 50, 250, and 1,000 highly expressed genes for predicting tissues from the ARCHS4 database after 10-fold cross validation. (D) Performance of a random forest classifier to predict ARCHS4 tissue type trained on GTEx top 250 GES or derived sGES.

samples. Specifically, we normalized and then averaged the reconstruction errors from autoencoders trained on GES and sGES (Fig. 6C and *SI Appendix, Fig. S14*). Compared to either the GES

(Fig. 6A) or sGES models (Fig. 6B), incorporating all signature information enables a clearer separation of true outliers in the data (Fig. 6C and *SI Appendix, Fig. S14*). For example, this

Fig. 6. Integrated signatures enable identification of robust signatures across databases. (A) Detection of outlier samples compared to GTEx gene signatures using a stacked denoising autoencoder trained to reconstruct gene signature membership from GTEx (green) gene signatures (of size 250 genes). Samples with high reconstruction error indicate that the sample is an outlier when compared to GTEx gene signatures. The red line indicates error values 2 SDs away from the mean of the distribution of errors reconstructing a validation GTEx set (error of 0.00725). (B) Outlier detection using distinct structural signature levels for muscle tissue. (C) GES and sGES autoencoders were combined by averaging the reconstruction error for each sample. Using this approach, we are able to identify the true outlier sample in ARCHS4, as compared to GTEx healthy tissue. (D) Predictive performance of GTEx GES to predict ARCHS4 tissue types, before and after outliers from ARCHS4 were removed. (E) Consistency of GES and sGES of across ARCHS4 and GTEx for muscle and whole-blood tissue types, before outlier removal (black) and after outlier removal (turquoise). Red line indicates a $J_c = 0.33$.

approach indicated that all pancreas tissue signatures from the ARCHS4 database can be considered outliers to GTEx pancreas signatures and thus validated that the pancreas signature is not robust across datasets. However, for tissues such as muscle, ovary, heart, and spleen, outliers can be easily identified (Fig. 6C). For instance, GSM1281783, the sample with the largest reconstruction error in heart tissue in ARCHS4, characterizes dilated cardiomyopathy. Likewise, GSM2071283 (muscle) represents a sample from fetal skeletal muscle tissue, which is different from healthy adult muscle cells characterized in GTEx (Fig. 6C).

Significant Improvement in GES and sGES Robustness after Outlier Removal. After outlier removal, we reevaluated the robustness of GES and sGES from GTEx using the ARCHS4 database. We observed a significant increase in predictive performance of GES after removing outliers (Fig. 6D and *SI Appendix, Fig. S15*) and a significant improvement in overlap between ARCHS4 and GTEx GES and sGES (Fig. 6E and *SI Appendix, Figs. S16 and S17*).

By removing outliers, we were able to identify specific GES and sGES that are common across all ARCHS4 and GTEx samples (*Dataset S1* and *SI Appendix, Table S7*). Specifically, for whole blood tissues, while only two genes are consistently seen across both datasets (actin beta and ferritin light chain), several domains (e.g., protein kinase domain and immunoglobulin-like domain), families (C1 set domains, pyruvate oxidase, and decarboxylase pyrophosphate-binding [PP] module), superfamilies (EF-hand and clathrin adaptor appendage domain), and folds (phosphate-binding loop [P-loop]-containing nucleoside triphosphate hydrolases and Src-homology 3 [SH3]-like barrel) were consistently observed, demonstrating that sGES can illuminate additional biological information not present in GES alone.

Taken together, our results indicate that integration of both gene and protein structure information (i.e., with GES and sGES, respectively) can dramatically improve the identification of outliers and enables the detection of robust expression signatures across datasets.

sGES Captures Drug Action on Cardiomyocyte-like Cell Lines. We investigated whether sGES alone can describe drug action on newly obtained transcriptomics data. We analyzed expression data from cardiomyocyte-like cell lines, generated by the Drug Toxicity Signature Generation (DToxS) LINCS Center, to identify perturbation-specific cardiomyocyte response to specific drugs (40). We observed that certain over- and underrepresented protein folds distinguish kinase inhibitor response from anthracycline drugs (Fig. 7A–C). For example, the kinase inhibitors nilotinib, regorafenib, sorafenib, pazopanib, and vemurafenib have a characteristic underexpression of folds relating to metabolism (*SI Appendix, Table S8*). Conversely, the anthracycline drugs epirubicin and doxorubicin have characteristic overexpression of folds related to cytokine action (*SI Appendix, Table S8*) as well as underexpression of folds relating to transfer RNA (tRNA) regulation such as: proline tRNA ligase; prolyl-tRNA synthetase; aminoacyl-tRNA synthetases; transmembrane ATPases; aminoacyl-tRNA synthetases; and anticodon-binding and contactin-binding protein (*SI Appendix, Table S9*). Taken together, fold-level sGES alone can further specify drug activity on cardiomyocytes, in addition to ranked lists of expressed genes.

Discussion

We hypothesized that transforming gene signature space into protein structure space can characterize a robust, reproducible signature and accurately define a phenotype. We also tested whether the integration of higher-order structural features (e.g., domain, fold, and superfamily), with ranked gene lists through a machine learning approach (i.e., autoencoder), can be used to improve reproducibility across transcriptomics data sets. Three key findings emerge from this study.

First, we developed a signature type, sGES, using features derived from multiple levels of protein structure (Fig. 1). sGES alone is able to characterize biological phenomena such as tissue type (Fig. 1). sGES also improved the overall consistency of GES while not impacting the predictive performance of signatures both within the same GES dataset as well as across gene expression datasets (i.e., GTEx and ARCHS4) (Figs. 3 and 4). The performance of sGES, as measured by the overlap of top-ranked genes based on expression level (J_C ; Fig. 3), is comparable across varying stringency cutoffs associating a gene to a structure. sGES derived from long alignments perform similarly to those based on short alignments (*SI Appendix, Fig. S18*).

This result may seem surprising because short alignments may correspond to local motifs that do not capture the “global” fold or function of proteins and thus would be expected to lead to lower J_C values among samples with distantly related gene sets. However, even short alignments or motifs can capture global protein properties; aligning just 11 residues that are 100% identical in sequence is highly predictive of protein fold (41). In fact, the ABC signature motif is a five-residue motif highly predictive of the ABC transporter architecture (42). Furthermore, shorter alignments [e.g., of five residues (42)] or alignments with similar similarity cutoffs (43) have been shown to capture distant functional and structural relationships in membrane protein families (43).

Second, we define complementary measures for evaluating the robustness of the GES; consistency corresponds to the overlap of top-ranked genes based on expression level (J_C ; Fig. 2). J_C is similar to previously proposed assessment measures of GES (10, 16), and predictivity assesses the predictive power of a phenotype using GES derived in different samples (Figs. 2–5). These measures assess different aspects of reproducibility in transcriptomics analysis. This result suggests that similar measures can be used in future studies analyzing gene expression patterns.

We also observed that integration of sGES and GES facilitates the identification of outliers among experimental samples as well as the interoperability between experiments that use differing transcriptomic methodologies improving the reproducibility of transcriptomics analysis studies (Fig. 6). We anticipate that integration of GES and sGES with other information such as protein–protein interaction networks can further improve the reproducibility of GES.

Third, sGES was benchmarked against independent approaches and datasets (Fig. 3D), demonstrating its general applicability. sGES enhanced the performance of commonly used databases and methods, including mSigDB (10) and the Human Protein Atlas (36). Adding an sGES representation on heterogeneous GES derived from these approaches improved the consistency of identifying related or unique phenotypes. sGES was also tested on an independent set of differentially expressed genes in response to prescription drugs (Fig. 7). sGES distinguishes kinase inhibitors from anthracyclines by identifying several folds related to tRNA regulatory factors that are down-regulated in response to anthracyclines. Doxorubicin and its analogs bind to tRNA molecules enabling their antitumor activity; however, down-regulation of tRNA regulatory factors has not been previously reported, demonstrating a potential mechanism of anthracycline action (44–46).

While sGES is a promising approach for analyzing transcriptomics data, sGES has one limitation. Principally, sGES cannot reliably calculate the enrichment of structural features for small gene lists or genes with limited structural annotation or diversity. Despite this limitation, the coming deluge in experimentally determined structures and new methods for predicting protein structure, such as AlphaFold (47), will increase the structural coverage of the human genome, allowing us to generate many more useful sGES. Therefore, we anticipate the sGES approach, which adds a new layer of information to gene expression patterns, will be helpful in defining robust identifiers of novel biological phenotypes.

Fig. 7. Characterization of kinase inhibitor activity using structural signatures. (A) t-SNE clustering of fold signatures from distinct type of drugs on Promocell cardiomyocyte-like cell lines. Rows are labeled by drug name or level 3 anatomical therapeutic chemical (ATC) category. (B) Overexpressed fold signatures for certain drugs. Distinct overexpressed clusters of folds are identified with numbers 1–5 and are described in *SI Appendix, Table S8*. (C) Underexpressed fold signatures for certain drugs. Distinct underexpressed clusters of folds are identified with numbers 1–4 and are described in *SI Appendix, Table S9*.

Materials and Methods

SI Appendix includes detailed information on the following: 1) computation of gene set consistency; 2) definition of an sGES; 3) a random forest algorithm for predicting tissue type; 4) predictivity of random forest models on

ARCHS4 gene sets; 5) sGES consistency; 6) benchmarking sGES against other methods for improving GES robustness; 7) clustering GTEx samples; 8) sGES predictivity; 9) integration of GES and sGES; 10) interoperability of GTEx and ARCHS4 GES; 11) experimental protocols for cell culture, drug treatment, and

transcriptomics; 12) classification of Promocell cardiomyocyte cell lines; 13) processing and exploratory analysis of gene expression data; 14) structural characterization of differentially expressed genes (DEGs) from perturbation studies; 15) and clustering of kinase inhibitors.

Data Availability. Data and code data have been deposited in Schlessinger Lab GitHub (<https://github.com/schlessinger-lab/structural-signatures>).

1. G. W. Gundersen *et al.*, GEN3VA: Aggregation and analysis of gene expression signatures from related studies. *BMC Bioinformatics* **17**, 461 (2016).
2. M. Li *et al.*, HiFreSP: A novel high-frequency sub-pathway mining approach to identify robust prognostic gene signatures. *Brief. Bioinform.* **21**, 1411–1424 (2020).
3. A. Shafi, T. Nguyen, A. Peyvandipour, S. Draghici, GSMA: An approach to identify robust global and test gene signatures using meta-analysis. *Bioinformatics* **36**, 487–495 (2020).
4. M. Ashburner *et al.*; The Gene Ontology Consortium, Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
5. The Gene Ontology Consortium, The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).
6. W. H. Khoo *et al.*, A niche-dependent myeloid transcriptome signature defines dormant myeloma cells. *Blood* **134**, 30–43 (2019).
7. M. C. Liu *et al.*, PAM50 gene signatures and breast cancer prognosis with adjuvant anthracycline- and taxane-based chemotherapy: Correlative analysis of C9741 (alliance). *NPJ Breast Cancer* **2**, 15023 (2016).
8. A. Subramanian *et al.*, A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452.e17 (2017).
9. A. Liberzon *et al.*, The molecular signatures database hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
10. A. Liberzon *et al.*, Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
11. J. Lonsdale *et al.*; GTEx Consortium, The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
12. A. Lachmann *et al.*, Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* **9**, 1366 (2018).
13. N. M. Ferraro *et al.*; TOPMed Lipids Working Group; GTEx Consortium, Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science* **369**, eaaz5900 (2020).
14. A. B. Keenan *et al.*, ChEA3: Transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Res.* **47**, W212–W224 (2019).
15. P. Patil, P.-O. Bachant-Winner, B. Haibe-Kains, J. T. Leek, Test set bias affects reproducibility of gene signatures. *Bioinformatics* **31**, 2318–2323 (2015).
16. L. Ein-Dor, O. Zuk, E. Domany, Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 5923–5928 (2006).
17. A. Raj, A. van Oudenaarden, Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell* **135**, 216–226 (2008).
18. K. Anderson *et al.*, Reproducibility of gene expression signature-based predictions in replicate experiments. *Clin. Cancer Res.* **12**, 1721–1727 (2006).
19. R. A. Ach *et al.*, Robust interlaboratory reproducibility of a gene expression signature measurement consistent with the needs of a new generation of diagnostic tools. *BMC Genomics* **8**, 148 (2007).
20. M. Crow, N. Lim, S. Ballouz, P. Pavlidis, J. Gillis, Predictability of human differential gene expression. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 6491–6500 (2019).
21. N. U. Rashid, Q. Li, J. J. Yeh, J. G. Ibrahim, Modeling between-study heterogeneity for improved reproducibility in gene signature selection and clinical prediction. *J. Am. Stat. Assoc.* **115**, 1125–1138 (2020).
22. T. E. Sweeney, W. A. Haynes, F. Vallania, J. P. Ioannidis, P. Khatri, Methods to increase reproducibility in differential gene expression via meta-analysis. *Nucleic Acids Res.* **45**, e1 (2017).
23. A. Dhawan *et al.*, Guidelines for using sigQC for systematic evaluation of gene signatures. *Nat. Protoc.* **14**, 1377–1400 (2019).
24. M. R. Birtwistle *et al.*, “Systems pharmacology: An overview” in *Systems Pharmacology and Pharmacodynamics*, D. E. Mager, H. H. C. Kimko, Eds. (Springer International Publishing, 2016), pp. 53–80.
25. S. Wang, E. R. Flynn, R. B. Altman, Gaussian embedding for large-scale gene set analysis. *Nat. Mach. Intell.* **2**, 387–395 (2020).
26. M. Rahilly, P. J. Carder, A. al Nafussi, D. J. Harrison, Distribution of glutathione S-transferase isoenzymes in human ovary. *J. Reprod. Fertil.* **93**, 303–311 (1991).
27. S. N. Kalam, S. Dowland, L. Lindsay, C. R. Murphy, Microtubules are reorganised and fragmented for uterine receptivity. *Cell Tissue Res.* **374**, 667–677 (2018).
28. B. Rost, J. Liu, R. Nair, K. O. Wrzeszczynski, Y. Ofra, Automatic prediction of protein function. *Cell. Mol. Life Sci.* **60**, 2637–2650 (2003).
29. Q. C. Zhang *et al.*, Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490**, 556–560 (2012).
30. X. Wang *et al.*, Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* **30**, 159–164 (2012).
31. L. Zimmermann *et al.*, A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J. Mol. Biol.* **430**, 2237–2243 (2018).
32. N. K. Fox, S. E. Brenner, J.-M. Chandonia, SCOPe: Structural classification of proteins—Extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **42**, D304–D309 (2014).
33. A. L. Mitchell *et al.*, InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, D351–D360 (2019).
34. L. van der Maaten, G. Hinton, Visualizing data using t-SNE. *J. Mach. Learning Res.* **9**, 2579–2605 (2008).
35. R. Dinga, B. W. J. H. Penninx, D. J. Veltman, L. Schmaal, A. F. Marquand, Beyond accuracy: Measures for assessing machine learning models, pitfalls and guidelines. *bioRxiv* [Preprint]. <https://doi.org/10.1101/743138> (22 August 2019).
36. M. Uhlen *et al.*, Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
37. O. Palasca, A. Santos, C. Stolte, J. Gorodkin, L. J. Jensen, TISSUES 2.0: An integrative web resource on mammalian tissue expression. *Database (Oxford)* **2018**, bay003 (2018).
38. T. Barrett *et al.*, NCBI GEO: Archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2013).
39. G. E. Hinton, R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507 (2006).
40. J. G. C. van Hasselt *et al.*, Transcriptomic profiling of human cardiac cells predicts protein kinase inhibitor-associated cardiotoxicity. *Nat. Commun.* **11**, 4809 (2020).
41. B. Rost, Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94 (1999).
42. P. C. Smith *et al.*, ATP binding to the motor domain from an ABC transporter drives formation of a nucleotide sandwich dimer. *Mol. Cell* **10**, 139–149 (2002).
43. A. Schlessinger *et al.*, Comparison of human solute carriers. *Protein Sci.* **19**, 412–428 (2010).
44. D. Agudelo, P. Bourassa, G. Bérubé, H. A. Tajmir-Riahi, Review on the binding of anticancer drug doxorubicin with DNA and tRNA: Structural models and antitumor activity. *J. Photochem. Photobiol. B* **158**, 274–279 (2016).
45. D. Agudelo, P. Bourassa, M. Beauregard, G. Bérubé, H.-A. Tajmir-Riahi, tRNA binding to antitumor drug doxorubicin and its analogue. *PLoS One* **8**, e69248 (2013).
46. S. Charak, M. Shandilya, R. Mehrotra, RNA targeting by an anthracycline drug: Spectroscopic and in silico evaluation of epirubicin interaction with tRNA. *J. Biomol. Struct. Dyn.* **38**, 1761–1771 (2019).
47. A. W. Senior *et al.*, Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).