Supplementary for "Gene expression inference with deep learning"

1 Removing duplicate profiles

The original Gene Expression Omnibus (GEO) [1] data consists of 129,158 gene expression profiles from the Affymetrix microarray platform. Some of the expression profiles are biological or technical replicates. To remove duplicated profiles, we clustered the 129,158 profiles into 100 clusters using k-means. Within each cluster, we calculated the pair-wise euclidean distance between each profile. We defined a pair of profiles as duplicates if their euclidean distance was less than 1.0, and removed one of the profiles. 18,149 duplicates were removed by this criterion, leaving 111,009 profiles remained in the GEO data.

2 Quantile normalization

We used three gene expression datasets generated by two different platforms in the study. The GEO data was generated by the Affymetrix microarray platform, with 22,268 probes corresponding to 978 landmark genes and 21,290 target genes. The 1000 Genomes (1000G) [2] expression data and the Genotype-Tissue Expression (GTEx) [3] data were generated by the Illumina RNA-Seq platform and measured based on Gencode V12 annotations [2, 3]. 943 probes out of the 978 landmark genes and 15,744 probes out of the 21,290 target genes have corresponding Gencode annotations in the RNA-Seq data. The 943 probes correspond to 943 Gencode annotations for the landmark genes. The 15,744 probes correspond to 9,520 Gencode annotations for the target genes as one Gencode annotation may include multiple probes. The original microarray data has been quantile normalized into a numerical range between 4 and 15 while the original RNA-Seq data are in the format of Reads Per Kilobase per Million (RPKM). To transfer the expression values of the RNA-Seq data into the same numerical scale as the microarray data while retaining the maximum information, we quantile normalized the three expression datasets together. Specifically, we used the following four steps to do quantile normalization:

1. We extracted the expression values of the 943 probes of landmark genes and the 15,744 probes of target genes from the microarray data, which have correspondence in Gencode annotations.

2. For those multiple probes in the microarray data that correspond to the same Gencode annotation in the RNA-Seq data, we took the mean of their expression values in the microarray data, resulted in 9,520 combined target genes that also have one-to-one correspondence between the two platforms.

3. The expression values of the 943 landmark genes and the 9,520 combined target genes from the microarray data were re-quantile-normalized together. A quantile of 10,463 genes was also generated.

4. For each profile in the RNA-Seq data, the expression values of the 943 landmark genes and the 9,520 target genes were ranked together and then mapped accordingly to the quantile computed in step 3.

3 Additional analysis

In the main manuscript, we used Mean Absolute Error (MAE) to measure the predictive performances of each method. To evaluate the performance of D-GEX and the other methods with different metrics, we also used Mean Squared Error (MSE) to evaluate the predictive performance at each target gene t,

$$\text{MSE}_{(t)} = \frac{1}{N'} \sum_{i=1}^{N'} \left(y_{i(t)} - \hat{y}_{i(t)} \right)^2 \tag{1}$$

where N' is the number of testing samples and $\hat{y}_{i(t)}$ is the predicted expression value for target gene t in sample i. The overall error is again defined as the average MSE over all target genes.

Additionally, we used the 1000 Genomes expression data as the validation dataset to monitor the training process for the GTEx expression data in the main manuscript. Since the 1000 Genomes expression data are all from the same lymphoblastoid cell line, which may have similar expression profiles, we also used a subset of the GTEx data as the validation dataset. Specifically, we randomly partitioned the GTEx data into ˜30% for validation (921 samples denoted as GTEx-sub-va) and ˜70% for testing (2,000 samples denoted as GTEx-sub-te).

We have redone all the experiments using MSE as the evaluation metric and GTEx-sub-va as the validation dataset. For the GEO expression data, D-GEX-10%-9000×3 achieves the best performance on both GEO-va and GEO-te. The relative improvements of D-GEX-10%-9000×3 are 24.42% over LR and 65.34 % over KNN-GE. D-GEX-10%-9000×3 outperforms LR in 99.98% of the target genes and outperforms KNN-GE in all the target genes. The complete performances of D-GEX with other dropout rates on both GEO-va and GEO-te are given in Supplementary Table S6 and S7. For the GTEx expression data, D-GEX-75%-9000×1 achieves the best performance on both GTEx-sub-va and GTEx-sub-te. The relative improvements of D-GEX-75%-9000×1 are 9.73% over LR and 47.24% over KNN-GE. D-GEX-75%-9000×1 outperforms LR in 84.46% of the target genes and outperforms KNN-GE in 94.52% of the target genes. The complete performances of D-GEX with other dropout rates on both GTEx-sub-va and GTEx-sub-te are given in Supplementary Table S8 and S9.

4 Supplementary figures

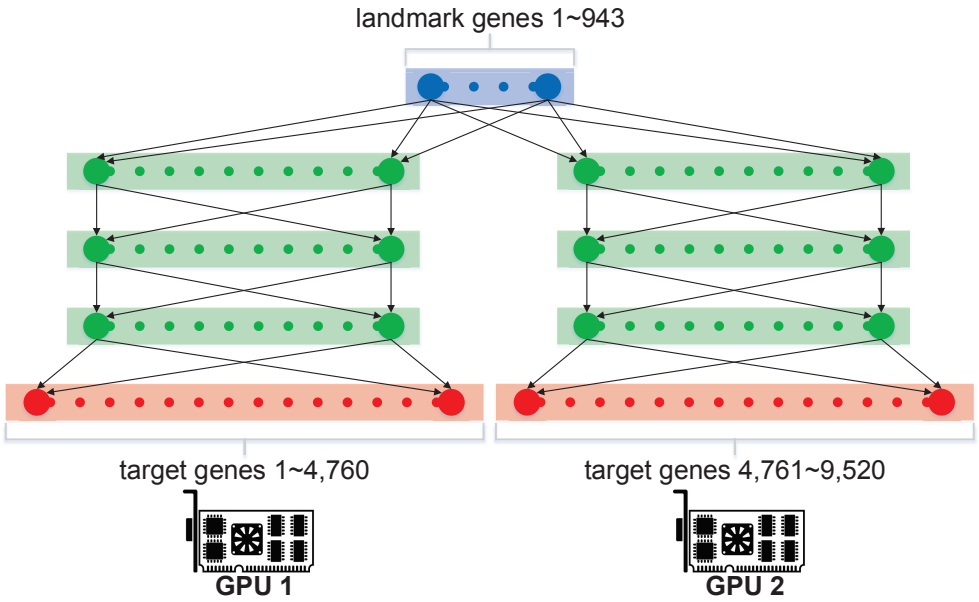

Figure S1: An example architecture of D-GEX with 3 hidden layers.

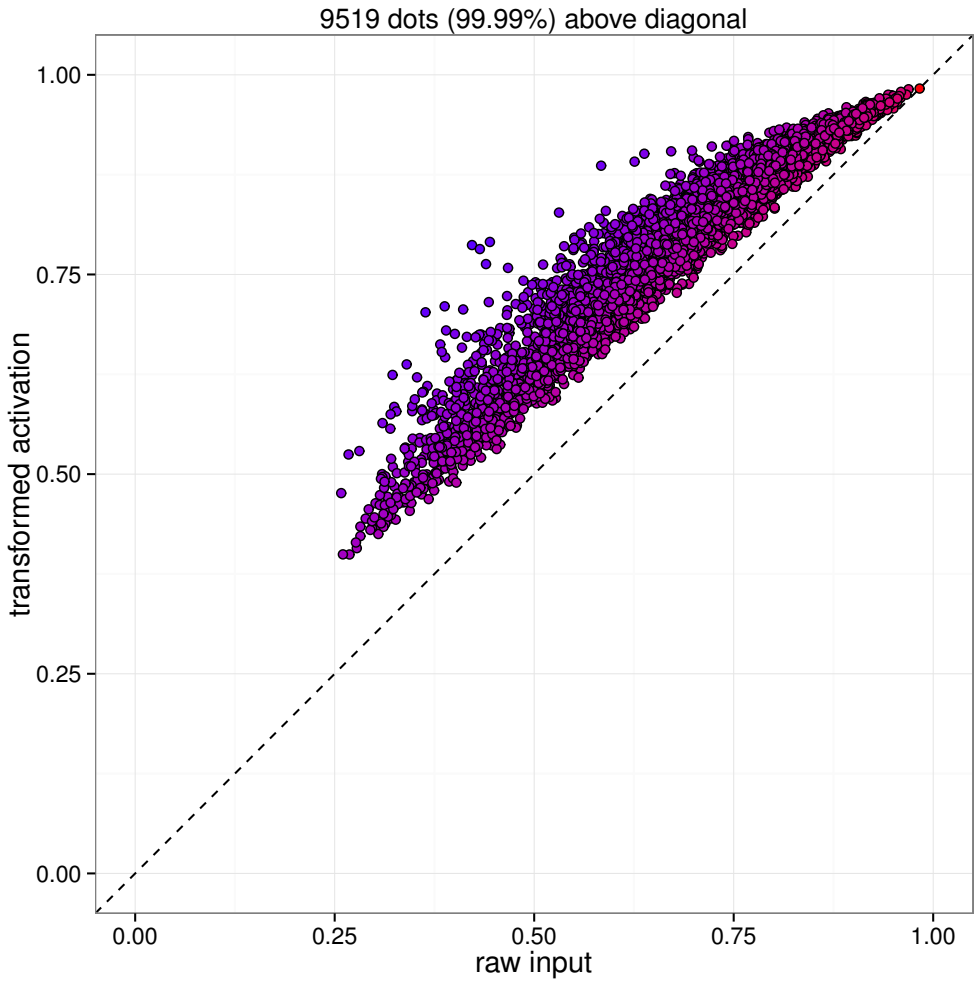

Figure S2: The adjusted R^2 of the transformed activations from D-GEX-10%-9000×3 compared to the adjusted R^2 of the raw inputs. Each dot represents 1 out of the 9,520 target genes. The x-axis is the adjusted R^2 of the raw inputs, and the y-axis is the adjusted R^2 of the transformed activations. Dots above diagonal means the transformed activations achieve larger adjusted R^2 compared to the raw inputs.

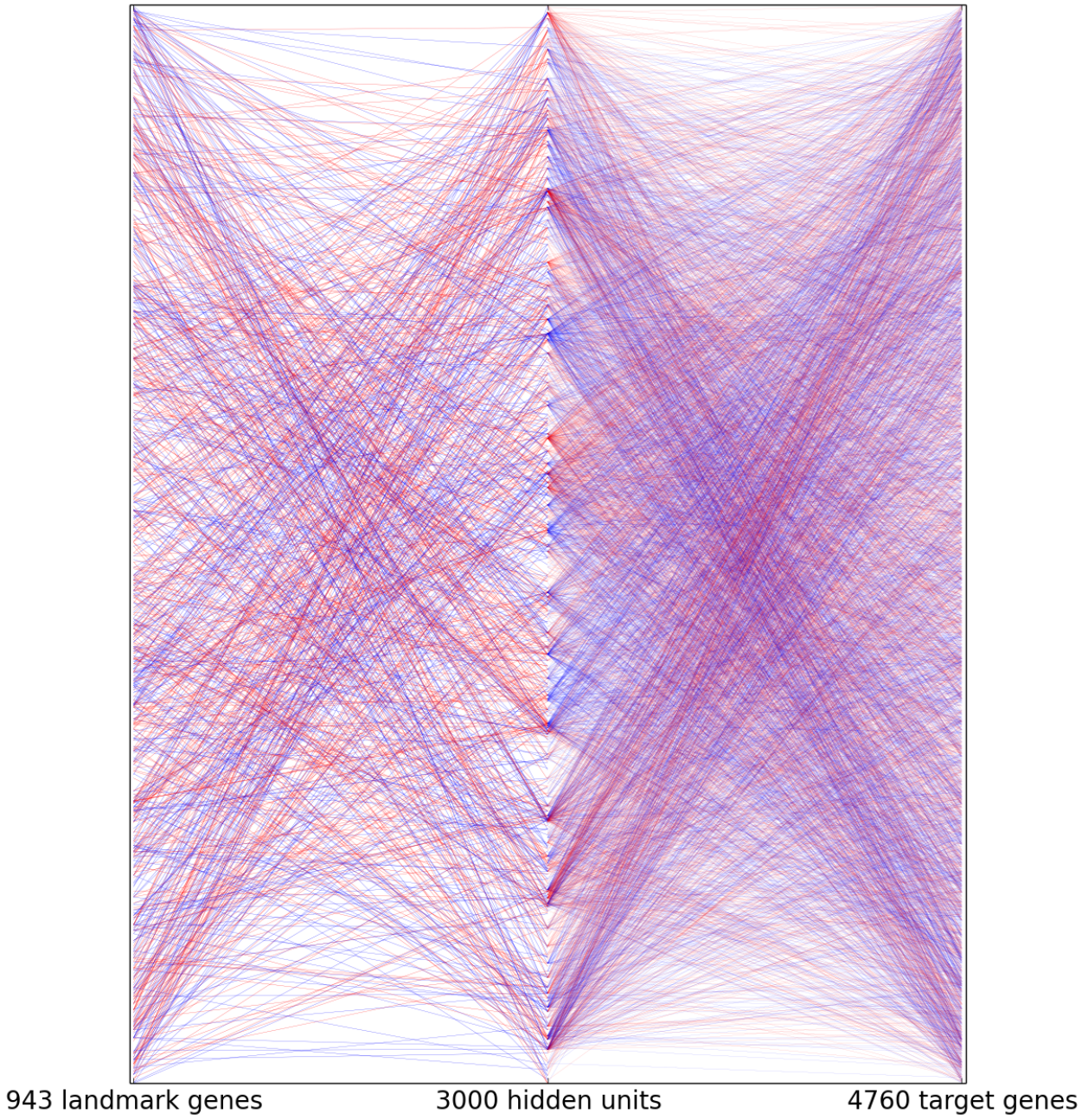

Figure S3: Visualizing the major weights of D-GEX-10%-3000×1 that was trained based on half of the target genes of GEO-tr and GEO-va. Red indicates postive weights and blue indicates negative weights.

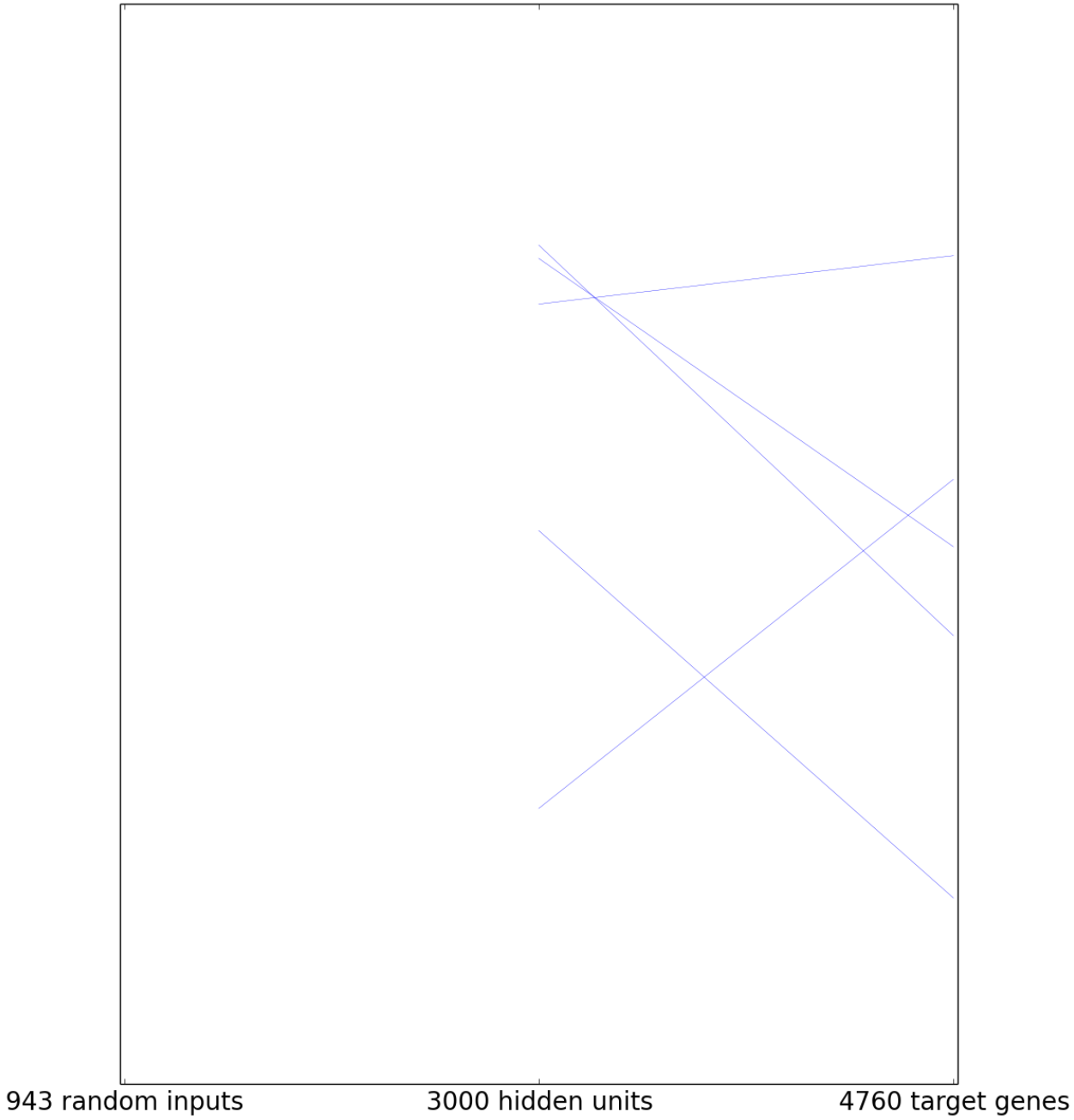

Figure S4: Visualizing the major weights of D-GEX-10%-3000×1 that was trained based on random inputs and half of the target genes of GEO-tr and GEO-va. Random inputs were sampled with the same dimension as the original inputs from normal distribution with 0 mean and 1 standard deviation for both training and validation. The model parameters were selected based on the performance on validation data. Red indicates postive weights and blue indicates negative weights.

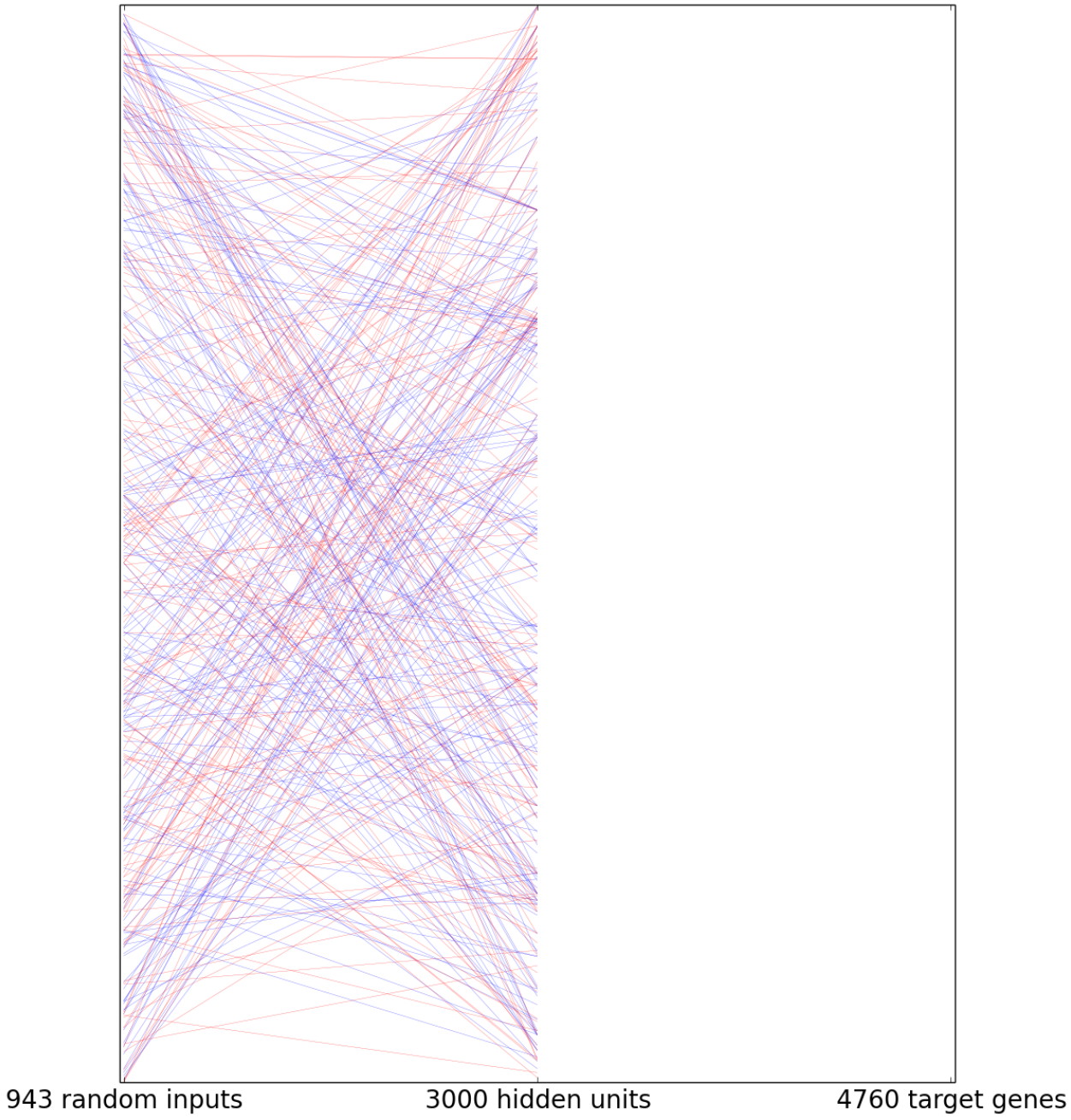

Figure S5: Visualizing the major weights of D-GEX-10%-3000×1 that was trained based on random inputs and half of the target genes of GEO-tr and GEO-va. Random inputs were sampled with the same dimension as the original inputs from normal distribution with 0 mean and 1 standard deviation for both training and validation. The model parameters were selected after 200 epochs of training. Red indicates postive weights and blue indicates negative weights.

5 Supplementary tables

Table S1: Detailed parameter configurations of D-GEX.

Parameters	
# of hidden layers	[1, 2, 3]
# of hidden units in each hidden layer	[3000, 6000, 9000]
Dropout rate	[0%, 10%, 25%]
Momentum coefficient	0.5
Initial learning rate [a]	5e-4 or 3e-4
Minimum learning rate	1e-5
Learning rate decay factor	0.9
Learning scale [b]	3.0
Mini-batch size	200
Training epoch	200
Weights initial range [c]	$\left[-\frac{\sqrt{6}}{\sqrt{n_i+n_o}}, \frac{\sqrt{6}}{\sqrt{n_i+n_o}}\right]$

[a] The initial learning rate is 5e-4 for architectures of 1 hidden layer and 3e-4 for architectures of 2 or 3 hidden layers.
[b] The learning rate of layers performing dropout was multiplied by a learning scale as suggested by [4].
[c] n_i, n_o denote the number of fan-ins and fan-outs of each unit. For the output layer with the linear activation function, the weights initial range is [-1e-4, 1e-4] instead.

Table S2: The MAE-based overall errors of LR, LR-L1, LR-L2, KNN-GE and D-GEX with different architectures and different dropout rates on GEO-va. Numerics after "±" are the standard deviations of prediction errors over all target genes. The best performance of D-GEX is underscored.

# of hidden units		3000	6000	9000
# of hidden layers and dropout rate	1, 25%	0.3503±0.0846	0.3425±0.0857	0.3399±0.0860
	2, 25%	0.3454±0.0841	0.3341±0.0855	0.3309±0.0861
	3, 25%	0.3471±0.0831	0.3323±0.0850	0.3265±0.0858
	1, 10%	0.3415±0.0854	0.3331±0.0865	0.3294±0.0870
	2, 10%	0.3370±0.0849	0.3273±0.0864	0.3217±0.0874
	3, 10%	0.3356±0.0845	0.3246±0.0864	<u>0.3197±0.0874</u>
	1, 0%	0.3421±0.0866	0.3336±0.0878	0.3294±0.0882
	2, 0%	0.3366±0.0866	0.3285±0.0884	0.3244±0.0893
	3, 0%	0.3339±0.0865	0.3265±0.0891	0.3233±0.0902
	LR		0.3776±0.0847	
	LR-L1		0.3774±0.0840	
	LR-L2		0.3776±0.0847	
	KNN-GE		0.5863±0.0697	

Table S3: The MAE-based overall errors of LR, LR-L1, LR-L2, KNN-GE and D-GEX with different architectures and different dropout rates on GEO-te. Numerics after "±" are the standard deviations of prediction errors over all target genes. The best performance of D-GEX is underscored.

# of hidden units		3000	6000	9000
# of hidden layers and dropout rate	1, 25%	0.3510±0.0851	0.3431±0.0861	0.3405±0.0864
	2, 25%	0.3461±0.0846	0.3348±0.0860	0.3315±0.0865
	3, 25%	0.3478±0.0837	0.3330±0.0855	0.3272±0.0863
	1, 10%	0.3421±0.0858	0.3337±0.0869	0.3300±0.0874
	2, 10%	0.3377±0.0854	0.3280±0.0869	0.3224±0.0879
	3, 10%	0.3362±0.0850	0.3252±0.0868	<u>0.3204±0.0879</u>
	1, 0%	0.3427±0.0871	0.3341±0.0882	0.3299±0.0886
	2, 0%	0.3372±0.0871	0.3291±0.0888	0.3251±0.0897
	3, 0%	0.3345±0.0870	0.3271±0.0895	0.3240±0.0907
	LR		0.3784±0.0851	
	LR-L1		0.3782±0.0844	
	LR-L2		0.3784±0.0851	
	KNN-GE		0.5866±0.0698	

Table S4: The MAE-based overall errors of LR, LR-L1, LR-L2, KNN-GE and D-GEX with different architectures and different dropout rates on 1000G-va. Numerics after "±" are the standard deviations of prediction errors over all target genes. The best performance of D-GEX is underscored.

# of hidden units		3000	6000	9000
# of hidden layers and dropout rate	1, 25%	0.7564±0.0810	0.7543±0.0861	0.7543±0.0886
	2, 25%	0.7578±0.0729	0.7486±0.0779	<u>0.7467±0.0811</u>
	3, 25%	0.7788±0.0635	0.7646±0.0740	0.7538±0.0768
	1, 10%	0.7598±0.0841	0.7563±0.0882	0.7540±0.0899
	2, 10%	0.7641±0.0717	0.7586±0.0824	0.7539±0.0839
	3, 10%	0.7748±0.0641	0.7710±0.0743	0.7645±0.0817
	1, 0%	0.7640±0.0714	0.7666±0.0761	0.7689±0.0784
	2, 0%	0.7648±0.0649	0.7647±0.0691	0.7657±0.0715
	3, 0%	0.7734±0.0646	0.7709±0.0686	0.7708±0.0710
	LR		0.8046±0.0977	
	LR-L1		0.7457±0.0911	
	LR-L2		0.8046±0.0977	
	KNN-GE		0.7467±0.0514	

Table S5: The MAE-based overall errors of LR, LR-L1, LR-L2, KNN-GE and D-GEX with different architectures and different dropout rates on GTEx-te. Numerics after "±" are the standard deviations of prediction errors over all target genes. The best performance of D-GEX is underscored.

# of hidden units		3000	6000	9000
# of hidden layers and dropout rate	1, 25%	0.4507±0.1231	0.4428±0.1246	0.4394±0.1253
	2, 25%	0.4586±0.1194	0.4446±0.1226	<u>0.4393±0.1239</u>
	3, 25%	0.5160±0.1157	0.4595±0.1186	0.4492±0.1211
	1, 10%	0.4518±0.1233	0.4450±0.1247	0.4399±0.1257
	2, 10%	0.4775±0.1190	0.4525±0.1221	0.4468±0.1239
	3, 10%	0.5069±0.1155	0.4784±0.1194	0.4561±0.1217
	1, 0%	0.4780±0.1199	0.4735±0.1213	0.4725±0.1219
	2, 0%	0.4934±0.1168	0.4822±0.1183	0.4790±0.1198
	3, 0%	0.5027±0.1151	0.4911±0.1173	0.4870±0.1190
	LR		0.4702±0.1234	
	LR-L1		0.5667±0.1271	
	LR-L2		0.4702±0.1234	
	KNN-GE		0.6520±0.0982	

Table S6: The MSE-based overall errors of LR, LR-L1, LR-L2, KNN-GE and D-GEX with different architectures and different dropout rates on GEO-va. Numerics after "±" are the standard deviations of prediction errors over all target genes. The best performance of D-GEX is underscored.

# of hidden units		3000	6000	9000
# of hidden layers and dropout rate	1, 25%	0.2614±0.1246	0.2519±0.1238	0.2487±0.1236
	2, 25%	0.2566±0.1230	0.2425±0.1219	0.2381±0.1218
	3, 25%	0.2593±0.1222	0.2408±0.1211	0.2336±0.1208
	1, 10%	0.2501±0.1236	0.2401±0.1230	0.2364±0.1229
	2, 10%	0.2443±0.1220	0.2336±0.1219	0.2280±0.1220
	3, 10%	0.2438±0.1212	0.2308±0.1212	<u>0.2255±0.1217</u>
	1, 0%	0.2508±0.1264	0.2416±0.1264	0.2372±0.1255
	2, 0%	0.2443±0.1252	0.2357±0.1262	0.2316±0.1261
	3, 0%	0.2412±0.1246	0.2342±0.1268	0.2307±0.1275
	LR		0.2984±0.1320	
	LR-L1		0.2980±0.1308	
	LR-L2		0.2984±0.1320	
	KNN-GE		0.6565±0.1542	

Table S7: The MSE-based overall errors of LR, LR-L1, LR-L2, KNN-GE and D-GEX with different architectures and different dropout rates on GEO-te. Numerics after "±" are the standard deviations of prediction errors over all target genes. The best performance of D-GEX is underscored.

# of hidden units		3000	6000	9000
# of hidden layers and dropout rate	1, 25%	0.2645±0.1270	0.2545±0.1259	0.2511±0.1255
	2, 25%	0.2599±0.1257	0.2452±0.1241	0.2407±0.1238
	3, 25%	0.2627±0.1252	0.2438±0.1235	0.2363±0.1229
	1, 10%	0.2529±0.1258	0.2425±0.1249	0.2388±0.1247
	2, 10%	0.2471±0.1243	0.2362±0.1240	0.2305±0.1239
	3, 10%	0.2468±0.1236	0.2334±0.1233	0.2281±0.1237
	1, 0%	0.2534±0.1284	0.2440±0.1283	0.2394±0.1272
	2, 0%	0.2467±0.1272	0.2382±0.1281	0.2340±0.1279
	3, 0%	0.2437±0.1266	0.2366±0.1286	0.2332±0.1294
	LR		0.3018±0.1341	
	LR-L1		0.3014±0.1331	
	LR-L2		0.3018±0.1341	
	KNN-GE		0.6582±0.1558	

Table S8: The MSE-based overall errors of LR, LR-L1, LR-L2, KNN-GE and D-GEX with different architectures and different dropout rates on GTEx-sub-va. Numerics after "±" are the standard deviations of prediction errors over all target genes. The best performance of D-GEX is underscored.

# of hidden units		3000	6000	9000
# of hidden layers and dropout rate	1, 25%	0.4035±0.2500	0.3924±0.2527	0.3874±0.2542
	2, 25%	0.4193±0.2440	0.3982±0.2476	0.3892±0.2494
	3, 25%	0.4488±0.2407	0.4196±0.2441	0.4041±0.2468
	1, 10%	0.4049±0.2524	0.3953±0.2546	0.3884±0.2561
	2, 10%	0.4189±0.2450	0.4069±0.2492	0.3987±0.2524
	3, 10%	0.4356±0.2430	0.4214±0.2482	0.4115±0.2501
	1, 0%	0.4393±0.2531	0.4276±0.2547	0.4187±0.2565
	2, 0%	0.4449±0.2477	0.4364±0.2518	0.4300±0.2540
	3, 0%	0.4526±0.2461	0.4456±0.2522	0.4407±0.2542
	LR		0.4291±0.2549	
	LR-L1		0.4231±0.2518	
	LR-L2		0.4291±0.2549	
	KNN-GE		0.7430±0.2631	

Table S9: The MSE-based overall errors of LR, LR-L1, LR-L2, KNN-GE and D-GEX with different architectures and different dropout rates on GTEx-sub-te. Numerics after "±" are the standard deviations of prediction errors over all target genes. The best performance of D-GEX is underscored.

# of hidden units		3000	6000	9000
# of hidden layers and dropout rate	1, 25%	0.3952±0.2316	0.3843±0.2331	0.3793±0.2340
	2, 25%	0.4103±0.2258	0.3895±0.2288	0.3807±0.2306
	3, 25%	0.4395±0.2233	0.4107±0.2259	0.3954±0.2286
	1, 10%	0.3964±0.2334	0.3869±0.2347	0.3799±0.2358
	2, 10%	0.4100±0.2264	0.3982±0.2305	0.3898±0.2334
	3, 10%	0.4264±0.2249	0.4119±0.2295	0.4023±0.2317
	1, 0%	0.4299±0.2343	0.4185±0.2355	0.4096±0.2361
	2, 0%	0.4364±0.2302	0.4275±0.2337	0.4204±0.2353
	3, 0%	0.4442±0.2286	0.4373±0.2347	0.4329±0.2370
	LR		0.4202±0.2366	
	LR-L1		0.4141±0.2338	
	LR-L2		0.4202±0.2366	
	KNN-GE		0.7189±0.2418	

References

[1] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[2] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter ACt Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, 2013.

[3] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

[4] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.