A generative adversarial network model for disease gene prediction with RNA-seq data

Xue Jiang[1,2,3], Wei Qian[1], Miao Chen[1], Liang Chen[1], Guan Ning Lin[1,2,3*],

1 Shanghai Mental Health Center, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China
2 Shanghai Key Laboratory of Psychotic Disorders, Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, Shanghai 200030, China
3 Brain Science and Technology Research Center, Shanghai Jiao Tong University, Shanghai 200030, China

¶Membership list can be found in the Acknowledgments section.
* nickgnlin@sjtu.edu.cn

Abstract

Deep learning models often need large amounts of training samples (thousands of training samples) to effectively extract hidden patterns in the data, achieving better results. However, in the field of bio-medicine, the omics data obtained by using advanced sequencing technology usually have few patient samples (tens to hundreds of samples). Due to the small sample problem, statistic methods and intelligent machine learning methods could not obtain convergent gene set when prioritizing biomarkers. Besides, mathematical models designed for prioritising biomarkers perform differently on different data set. Nevertheless, the architecture of generative adversarial network (GAN), which is good at unsupervised learning, can address the bottleneck problem. Generator in GAN could generate large amounts of samples with similar distributions to that of samples in training set, thus improving prediction accuracy and robustness of the discriminator. So, in this study, we designed a new generative adversarial network model with variational auto-encoder (VAE) as generator and multi-layer perceptron (MLP) as discriminator. VAE could capture hierarchical structure and probability distribution of input data. The prediction residual was backpropagated to the generator part of VAE, modifing the captured probability distribution. Through the game between VAE and MLP, the prediction performance could be further improved. Based on this model, we further designed a disease gene prediction framework to predict disease genes with RNA-seq data. Experimental results highlight key genes and pathways underlying complex disease phenotypes.

Author summary

Our deep learning model improves the identification accuracy of disease genes over the state of the art approaches and can be extended to other omics data analysis. Our experimental analysis have uncovered new disease-related genes and disease-associated pathways in the brain, which in turn have provided insight into the molecular mechanisms underlying disease phenotypes.

Introduction

With the rapid development and decreasing cost of sequencing technologies, large amounts of omics data become available, providing not only opportunities but also challenges for decoding the pathophysiologic mechanism of chronic complex diseases from a system molecular level. It is an important research topic in the field of bioinformatics that develop efficient and reliable computational tools to screen disease biomarkers and drug targets with omics data.

The development of machine learning algorithms in artificial intelligence applications effectively promote the omics data mining and analyzing, accelerating disease-related biomarker and pathway discovery, as well as pushing the discovery of regularity and specificity under disease phenotypes. Machine learning methods can be classified into two categories. One is generative model, such as Gaussian mixture model [1], Hidden Markov model [2], Restrict Boltzmann machine [3], deep belief network [4], and auto-encoder [5], etc. The other category is discriminative model, such as linear regression model [6], linear discriminant analysis [7, 8], support vector machine [9], and multi-layer perception [10], etc. The generative model learns the potential spatial distribution and characterization of data, evaluates joint probability density distribution from the data, and finally solves conditional probability as the prediction model [11]. The discriminative model learns decision function or conditional probability distribution. Due to discriminative model conduct prediction directly, it often achieves higher accuracy and could conduct various abstract to the data.

Recent advance in deep learning demonstrate promising performance in various fields, including protein structure prediction, natural image analysis, and natural language processing. Deep learning methods often need large amounts of training samples (thousands of training samples) to effectively extract hidden patterns in the data, achieving better results. In the field of biomedicine, biomedical data often have few training samples (tens to hundreds of samples) due to high cost of biochemical experiments and long time consuming. Small sample problem makes it difficult to screen robust disease biomarkers with omics data, resulting in poor reproducible of prediction results among different patients. The above disadvantage greatly limits the application of supervised machine learning and deep learning methods. Meanwhile, it is also difficult to design a suitable loss function to learn the potential distribution of the training data. Nevertheless, the architecture of generative adversarial network (GAN), which is good at unsupervised learning, can address the bottleneck problem. Generator in GAN could generate large amounts of samples with similar distributions to that of samples in training set. Through the game between the unsupervised generator and the supervised discriminator, data that fit the distributions of training samples are generated by the generative model. With increasing of the generated data, the prediction performance of discriminator could be greatly improved [20, 21]. Therefore, the architecture of GAN could be used to solve the bottleneck problem of small samples in the field of biomedicine, improving the prediction accuracy and robustness of clinically useful biomarkers.

The generative adversarial network is a deep learning framework, which first puts generative model and discriminative model into one learning framework [12]. The two models are alternately iteratively training synchronously, and the training process is accomplished when the two models convergent to a Nash equilibrium [13]. During training process, the discriminative model guides the generative model to learn the probability density distribution of training samples, while the false data generated by the generative model help to improve the classification performance of the discriminative model [14–16]. Theoretically, GAN can train any kinds of generators, while other frameworks need the generator satisfy some specific function form, such as the output layer is Gaussian distribution. Compared with other generative deep learning models, the gradients are introduced from the discriminator, thus, the generator doesn't need to be updated according to the training samples. Besides, GAN does not need to repeatedly sample with Markov chain, or conduct inference during learning process. Due to various advantages of GAN, it has a wide applications in image editing [17], data generation [18], and attention prediction [19], etc.

Based on above discussion, to solve the small training sample problem in the field of bio-medicine, we designed a new generative adversarial network model with variational auto-encoder as generator and multi-layer perceptron as discriminator (GAN-VAEMLP). Multi-layer perceptron (MLP) with more than three layers can fit arbitrarily complex functions, and learn the mapping relationship well between the input data and output labels. Meanwhile, the variational auto-encoder (VAE) combines unsupervised learning with the variational Bayesian method perfectly [22]. VAE could capture hierarchical structure and probability distribution in the high-dimensional space of the input data, by applying probability transformation to basic auto-encoder which takes the input, hidden representation and reconstructed output as random probability variables in a directed graph. The prediction residual was backpropagated to the generator part of VAE, modifing the captured probability distribution. Through the game between VAE and MLP, the prediction performance could be further improved. Based on this model, we further designed a disease gene prediction framework to predict disease genes with RNA-seq data. Finally, to verify the disease gene prediction performance of GAN-VAEMLP, we also conduct experiments with DESeq2 [23], edgeR [24], limma [25], t-test [26], fold change method (FC) [26], GAN, and MLP. Comparison results shown a better performance of GAN-VAEMLP. Through the intergrated of top disease-associated genes in the 8 gene rank list obtained by the 8 approaches, we selected 9 key genes finally. GO and KEGG functional annotation of the 9 genes shown that biological processes, such as calcineurin-NFAT signaling cascade, mitotic cell cycle phase transition, and protein dephosphorylation, are seriously affected in the brain tissues of Huntington's disease mice.

The rest of this paper is organized as follows: In Section 2, we present the proposed GAN-VAEMLP model in detail. In Section 3, we illustrate experiments of different methods with RNA-seq data of Huntington's disease. The function annotation and pathways involved in the disease are analyzed and reported. And the overall discussion of experimental results of various methods are also reported. In Section 4, conclusions are presented.

Materials and methods

Gene expression data

We downloaded gene expression data from http://www.hdinhd.org, which are obtained from the striatum tissues of Huntington's disease mice through RNA-seq technology. The age of experimental mice include 2-month-old, 6-month-old, and 10-month-old. There are six genotypes for gene expression data at each stage, including poly Q20, poly Q80, poly Q92, poly Q110, poly Q140, and poly Q170. The number of samples for each genotype is illustrated in Table 1. Totally, there are 209 samples, including 48 samples of poly Q20, 32 samples for each other genotype, respectively. The modifier genes are from literature [29, 30], including 89 disease-associated genes, and 431 non-disease-associated genes.

Huntington's disease is a kind of hereditary neurodegenerative disease. The symptoms of the disease become more and more serious as the disease progress, including uncontrollable dance movements, mental disorders, personality changes, and mental deficiency. Once these symptoms appear, patients can only survival 10 to 15 years. The disease is caused by a triplet (CAG) repeat elongation in huntingtin (HTT) gene on chromosome 4 that codes for polyglutamine in the huntingtin protein [31]. The normal gene contains 11 to 28 CAG repeats. Individual with 29 to 34 repeats are almost impossible to develop the disease, while individual with 35 to 41 repeats may develop the disease with mild symptoms. The more CAG repeats, the earlier the disease occur and the more serious the symptoms are.

Table 1. Experimental data description.

Age	2-month-old	6-month-old	10-month-old			
Tissue	Striatum	Cortex	Liver			
Genotype	poly Q20	poly Q80	poly Q92	poly Q111	poly Q140	poly Q175

Discriminator

The perceptron is a linear model for dichotomy, which cannot classify nonlinear data. Theoretically, three-layer network can fit arbitrarily complex functions. Therefore, we deepen the hierarchy of the network, and used a three layer perceptron as the discriminator. Forward propagation was used to transfer information from the input layer to output layer, and parameters in the networks were updated using gradient back propagation algorithm.

Let $X = [x_1, \ldots, x_n]$ represent sample, and L represent the corresponding label. L_l represent the neurons in $l-th$ layer. y_l represent the outputs of $l-th$ layer, and y_l^j represent the output of $j-th$ neuron in $l-th$ layer. u_l represent the iutputs of $l-th$ layer, and u_l^j represent the input of $j-th$ neuron in $l-th$ layer. W_l represent the weighted connection between $l-th$ layer and $l-1-th$ layer. w_l^{ji} represents the weight between $i-th$ neuron in $l-1-th$ layer and $j-th$ neuron in $l-th$ layer. b_l represent bias of neurons in $l-th$ layer, and b_l^j represent the bias of $j-th$ neuron in $l-th$ layer. $\Theta = (W; b)$ represents the parameters in the network.

Forward propagation was used to transfer information from low layer to high layer, therefore

$$y_l^j = f(u_l^j). \tag{1}$$

$$u_l^j = \sum_{l \in L_{l-1}} (w_l^{ji} y_{l-1}^i + b_l). \tag{2}$$

$$y_l = f(u_l) = f(W_l y_{l-1} + b_l). \tag{3}$$

$f(.)$ is activation function. In this study, ReLU was used as activation function. We used mean square error as loss function.

$$E = \frac{1}{2} \sum_{i=1}^{n} (y_i - L_i)^2. \tag{4}$$

To minimize the loss function, back propagation algorithm and gradient descent was used to update parameters in the network.

$$\begin{aligned}
\frac{\partial E}{\partial y_l^j} =& \frac{\partial E(u_{l+1}^1, \dots, u_{l+1}^m)}{\partial y_l^j} \\
=& \sum_{k \in L_{l+1}} \frac{\partial E}{\partial u_{l+1}^k} \frac{\partial u_{l+1}^k}{\partial y_l^j} \\
=& \sum_{k \in L_{l+1}} \frac{\partial E}{\partial y_{l+1}^k} \frac{\partial y_{l+1}^k}{\partial u_{l+1}^k} \frac{\partial u_{l+1}^k}{\partial y_l^j}. \\
=& \sum_{k \in L_{l+1}} \frac{\partial E}{\partial y_{l+1}^k} \frac{\partial y_{l+1}^k}{\partial u_{l+1}^k} w_{l+1}^{kj}
\end{aligned} \tag{5}$$

Here, we defined the sensitivity of the node as the change rate of the error to the input,

$$\delta = \frac{\partial E}{\partial u}. \tag{6}$$

The sensitivity of the $j-th$ node in $l-th$ layer is

$$\delta_l^j = \frac{\partial E}{\partial u_l^j} = \frac{\partial E}{\partial y_l^j} \frac{\partial y_l^j}{\partial u_l^j} = \frac{\partial E}{\partial y_l^j} f'(u_l^j). \tag{7}$$

So,

$$\frac{\partial E}{\partial y_l^j} = \sum_{k \in L_{l+1}} \frac{\partial E}{\partial y_{l+1}^k} \frac{\partial y_{l+1}^k}{\partial u_{l+1}^k} w_{l+1}^{kj} = \sum_{k \in L_{l+1}} \delta_{l+1}^k w_{l+1}^{kj}. \tag{8}$$

multiply $f'(u_l^j)$ in both side of the Eq. 8, we get

$$\delta_l^j = \frac{\partial E}{\partial y_l^j} f'(u_l^j) = f'(u_l^j) \sum_{k \in L_{l+1}} \delta_{l+1}^k w_{l+1}^{kj}. \tag{9}$$

Then, we use gradient descent to update parameters

$$\frac{\partial E}{\partial w_l^{ji}} = \frac{\partial E}{\partial y_l^j} \frac{\partial y_l^j}{\partial u_l^j} \frac{\partial u_l^j}{\partial w_l^{ji}} = \frac{\partial E}{\partial y_l^j} f'(u_l^j) y_{l-1}^i. \tag{10}$$

$$\frac{\partial E}{\partial b_l^j} = \frac{\partial E}{\partial y_l^j} \frac{\partial y_l^j}{\partial u_l^j} \frac{\partial u_l^j}{\partial b_l^j} = \frac{\partial E}{\partial y_l^j} f'(u_l^j). \tag{11}$$

For the output layer,

$$\theta_l^j = \frac{\partial E}{\partial y_l^j} f'(u_l^j) = f'(u_l^j)(y_l^j - L^j). \tag{12}$$

$$\theta_l = (y_l - L) f'(u_l). \tag{13}$$

For the hidden layer,

$$\theta_l^j = \frac{\partial E}{\partial y_l^j} f'(u_l^j) = f'(u_l^j) \sum_{k \in L_{l+1}} \theta_{l+1}^k w_{l+1}^{kj}. \tag{14}$$

$$\theta_l = (W_{l+1}^T \theta_{l+1}) f'(u_l). \tag{15}$$

The matrix form can be written as

$$\frac{\partial E}{\partial W_l} = \theta_l y_{l-1}^T. \tag{16}$$

$$\frac{\partial E}{\partial b_l} = \theta_l. \tag{17}$$

Then, we can update parameters as follows:

$$W_l = W_l - \eta \frac{\partial E}{\partial W_l} = W_l - \eta \theta_l y_{l-1}^T. \tag{18}$$

$$b_l = b_l - \eta \frac{\partial E}{\partial b} = b_l - \eta \theta_l. \tag{19}$$

η is learning rate, and $\eta \in (0, 1)$. When η is large, the training process can convergent quickly. When η is small, the training process convergent slowly with a high precision.

Generator

To accurately capture the intrinsic features of original data and reconstruct it, we used variational auto-encoder (VAE) as the generator [32, 33]. In VAE, the encoder is a variational inference network that maps the observed input to the posterior distribution of the potential space. And the decoder is a generation network that maps any potential coordinates back to the distribution of the original data space.

We added a constraint to the encoder network to make sure the result roughly follows the unit Gaussian distribution, and then reconstruct the original data by transmitting the latent vector to the decoder network.

Let $X = [x_1, \cdots, x_n]$ represents sample, and $z = [z_1, \cdots, z_m]$ represents latent variables. $p(x)$ represents the distribution of sample, and $p(z)$ represents the probability distribution of latent variable. $p(x|z)$ is the probability of data generated of a given latent variable.

We assume that there exist latent variables z, which could generate an observation x. However, we want to inference the feature of z from the observed variable x, i.e., we need to compute $p(z|x)$.

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}. \tag{20}$$

Since $p(x) = \int p(x|z)p(z)dz$ is a complicate distribution and difficult to compute. We apply variational inference to estimate it. That is to say, we use the distribution $q(z|x)$ to approximate $p(z|x)$. Kullback-Leibler divergence can measure the difference between two probability distribution. Thus, we use Kullback-Leibler divergence to ensure the approximate between $q(z|x)$ and $p(z|x)$.

$$minKL(q(z|x)||p(z|x)). \tag{21}$$

It is equivalent to minimize the following equation,

$$L(\theta) = -E_{z\ q_\theta(z|x)}[logp(x|z)] + KL(q_\theta(z|x)||p(z)). \tag{22}$$

The first term is reconstruction loss denoted by the expectation of negative log likelihood, which encourages the model to be sensitive to the input and help the decoder to reconstruct the data well. The second term represents the Kullback-Leibler divergence between the distribution of $q(z|x)$ and $p(z)$. When the output of encoder z is different from normal distribution, it acts as a regularized factor and presents overfitting. A scaling parameter is usually added before the regularization to adjust the balance between the two targets. Then, the loss function of VAE can be written as below

$$L(\theta) = -E_{z\ q_{\theta}(z|x)}[logp(x|z)] + \beta KL(q_{\theta}(z|x)||p(z)). \tag{23}$$

Variational auto-encoder can learn a smooth potential representation of the input data. If we observe a dense distribution in latent space, we need to give larger weight to the KL divergence term, i.e. $\beta > 1$, encouraging the network learning a more wider distribution.

The error back propagation was used to update the parameters in the encoder and the decoder. Reparameterization trick method was used to randomly sample from unit. With this reparameterization, the parameters of the distribution can be optimized, maintaining the ability to randomly sample from the distribution [33, 34].

The generative adversarial network model

In this study, we use the architecture of GAN to solve small sample problem in the field of bio-medicine. Based on the GAN-VAEMLP, we desigend a framework to predicted disease-associated genes. The details are illustrated in Fig. 1. Finally, we could rank the disease-associated genes in descending order according to their score. Top ranking genes are most likely to be disease genes. The loss function of MLP is denoted as D, and that of VAE is denoted as G.

In the GAN-VAEMLP, The loss function for the discriminator is shown below:

$$L_D = -E_{x\ p_{data}(x)}[logD(x)] - E_{x\ p_{data}(x)}[log(1 - D(G(x)))]. \tag{24}$$

The loss function for the generator is shown below:

$$L_G = -E_{x\ p_{data}(x)}[log(D(G(x)))]. \tag{25}$$

Training

Since the freedom degree of GAN is too high, it should be well synchronized between discriminator and generator, which is hard to balance in actual training process. We use an alternate iterative strategy to train VAE and MLP. Besides, during the training process, discriminator is easily to convergent, while the generator is often divergent. To avoid pattern loss and accelerate the convergent speed, we make the input of both VAE and MLP to be samples in the training set. Until the VAE learns the intrinsic features of the training data, the performance of both the generator and the discriminator are all optimized. Then, the trained discriminator is used to predict the labels of unlabeled samples with gene expression data.

The prediction residual of MLP was backpropagated to the generator part of VAE, modifing the captured probability distribution of input data for VAE. Mini-batch random gradient descent training was used to train the networks in both generator and discriminator.

In summary, the detailed training process of the model is shown below.

High parameter setting: $k = 1, \beta = 2$.

GAN-VAEMLP.png

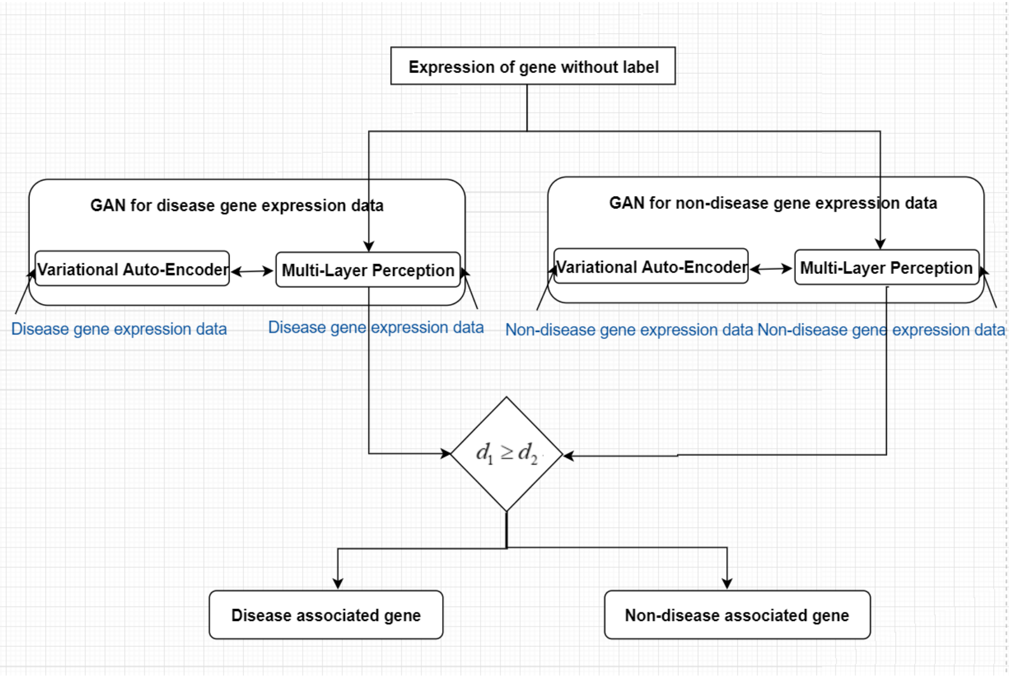

Expression of gene without label
GAN for disease gene expression data
Variational Auto-Encoder
Multi-Layer Perception
GAN for non-disease gene expression data
Variational Auto-Encoder
Multi-Layer Perception
Disease gene expression data
Disease gene expression data
Non-disease gene expression data
Non-disease gene expression data
$d_1 \geq d_2$
Disease associated gene
Non-disease associated gene

Fig 1. The Flowchart of GAN-VAEMLP for disease gene prediction.

```
Algorithm 1: Training for GAN-VAEMLP
For number of training iterations do
For $k$ step do
    Sample mini-batch of samples from original data prior
    Sample mini-batch of samples from data generating distribution
    update the MLP by descending stochastic gradient of Eq.23
End for
Update the decoder part of VAE using gradients introduced from MLP
Sample mini-batch of samples from data prior
Update the VAE by descending stochastic gradient of Eq. 24
End for
```

Results

Performance of GAN-VAEMLP

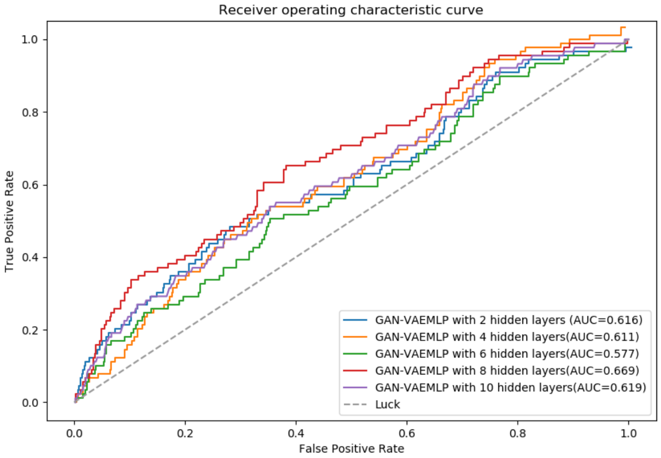
To verify the performance of GAN-VAEMLP, we conducted large amounts of experiments with different parameters. Finally, the training parameters for generator and discriminator are decided, which are shown in Table 2 and Table 3 respectively. In addition, we test the prediction accuracy of GAN-VAEMLP with different depth of VAE. The comparative results are shown in Fig. 2 and Fig. 4. We can well known that GAN-VAEMLP with 8 hidden layers in VAE performs best. So, we used the prediction results of this model to conduct comparison with other the state of the art approaches.

Table 2. Model training parameters for VAE.

Item	Value
Activation function for hidden layer	Relu
Dropout	0.25
Training batch size	64
Training epochs	500
Optimizer	Adam
Learning rate	0.001
Decay	0.9
Loss function	Binary_crossentropy + $\beta KLdivergence$

Table 3. Model training parameters for MLP.

Item	Value
Activation function for hidden layer	Relu
Activation function for output layer	sigmoid
Training batch size	64
Training epochs	200
Optimizer	Adam
Learning rate	0.001
Decay	0.9
Loss function	Mean square error

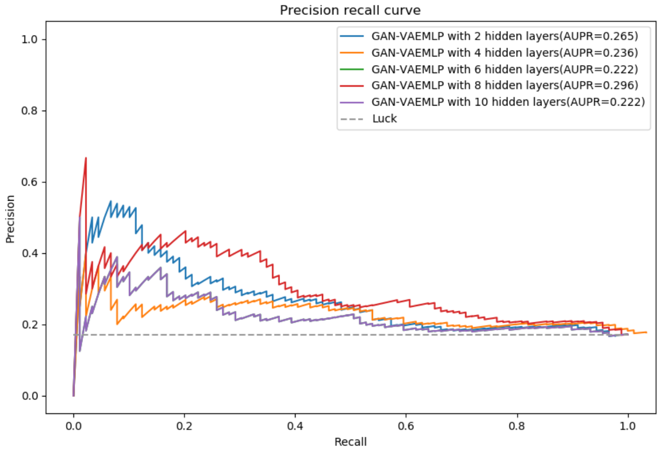


ROC.png

Fig 2. The ROCs of GAN-VAEMLP with different depth of VAE.

Comparison of GAN-VAEMLP with other state of the art approaches

To investigate the effectiveness of GAN-VAEMLP proposed in this study, we also conducted experiments with t-test, fold change method (FC), DESeq2, edgeR, limma, MLP, and GAN. The comparison results of the 8 methods are shown in fig. 4 and fig. 5. It is clearly shown that GAN-VAEMLP performs best compared with other approaches.



PR.png

Fig 3. The precision recall curves of GAN-VAEMLP with different depth of VAE.

There is a high prediction accuracy for top ranking genes for the 8 methods. Therefore, we get a robust differentially expressed gene set by intersecting top ranking 1000 genes in the ranking lists obtained by the 8 approaches, which are intuitively considered to be related with the disease. Finally, Traip, Bsgnt2, Ugt8a, Ppp3ca, Pmepa1, Rgs4, Ppp3r1, Chn1, and St8sia3 were selected out. We draw a cluster heatmap of the 9 feature genes using the expression under all samples, see Fig. 6. Obviously, samples of ploy Q20 can be clustered using the 9 feature genes, and the 9 feature genes can be clustered in one category.

To computational verify the effectiveness of the 9 biomarkers, we conducted genotype classification based on support vector machine. The classification accuracy of control genotype (poly Q20) from case genotype (poly Q ¿ 32) by using disease genes in the training set is 0.759, while the classification accuracy is 0.827 by using the 9 feature genes. It indicates that the expression of the 9 genes have been changed obviously with the repeat elongation in huntingtin gene.

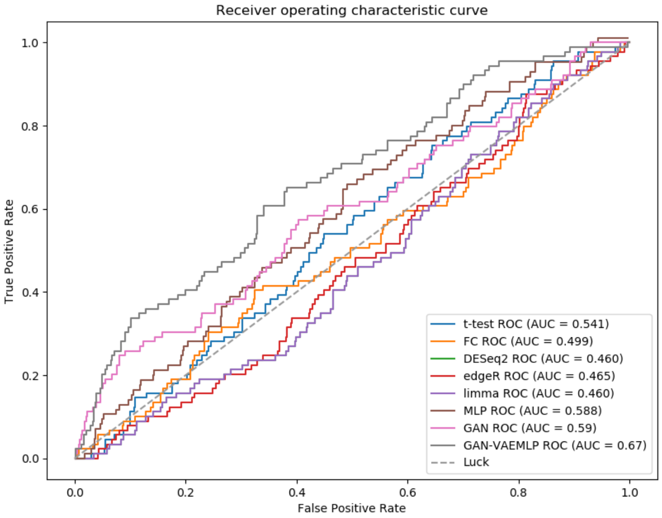
To get a deep insight into the dynamic molecular mechanism and pathological mechanism underlying complicated clinical disease phenotype, we performed gene functional analysis, enrichment analysis, and protein-protein interaction network analysis.

Gene functional analysis

We conducted gene functional analysis of the 9 genes using PsyMuKB (http://www.psymukb.net/). Meanwhile, there is no record for Bsgnt2 and Ugt8a. The other 7 genes are all protein coding genes. The detailed descriptions for the 7 feature genes are shown below.

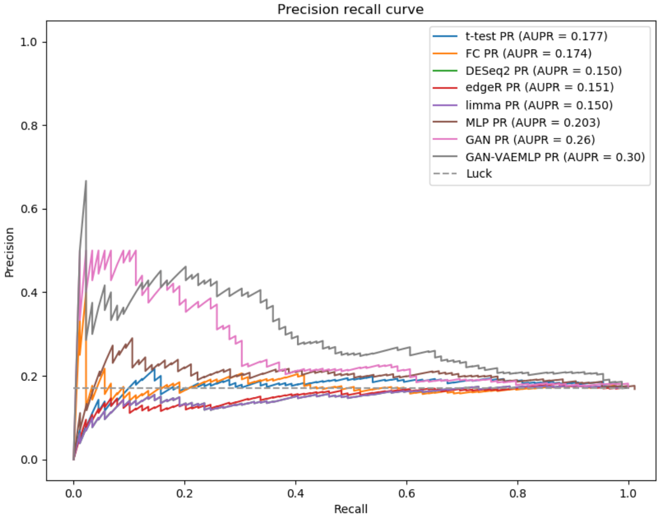
Traip (TRAF Interacting Protein) locate at 3p21.31. It has been reported that Seckel Syndrome 9 and Seckel Syndrome are associated with Traip. Gene ontology annotations related to this gene include ligase activity and obsolete signal transducer activity, downstream of receptor.

Ppp3ca (protein phosphatase 3 catalytic subunit alpha) locate at 4q24. Epileptic encephalopathy, infantile or early childhood, 1 and arthogryposis, cleft palate, craniosynostosis, and impaired intellectual development are associated with the gene. Gene ontology annotations for the gene are calcium ion binding and enzyme binding.



ROC.png

Fig 4. The ROCs of t-test, FC, DESeq2, edgeR, limma, MLP, GAN, and GAN-VAEMLP.



PR.png

Fig 5. The precision-recall curves of t-test, FC, DESeq2, edgeR, limma, MLP, GAN, and GAN-VAEMLP.

Pmepa1 (Prostate Transmembrane Protein) locate at 20q13.31. Gene ontology annotations are WW domain binding and R-SMAD binding.

Rgs4 (regulator of G protein signaling 4) locate at 1q23.3. Schizophrenia and psychotic disorder are associated with Rgs4. Gene ontology annotations include GTPase activity and G-protein alpha-subunit binding.

Ppp3r1 (Protein Phosphatase 3 Regulatory Subunit B, Alpha) locate at 2p14. Diseases associated with PPP3R1 include extracranial neuroblastoma and cervical neuroblastoma. Gene ontology annotations related to this gene include calcium ion

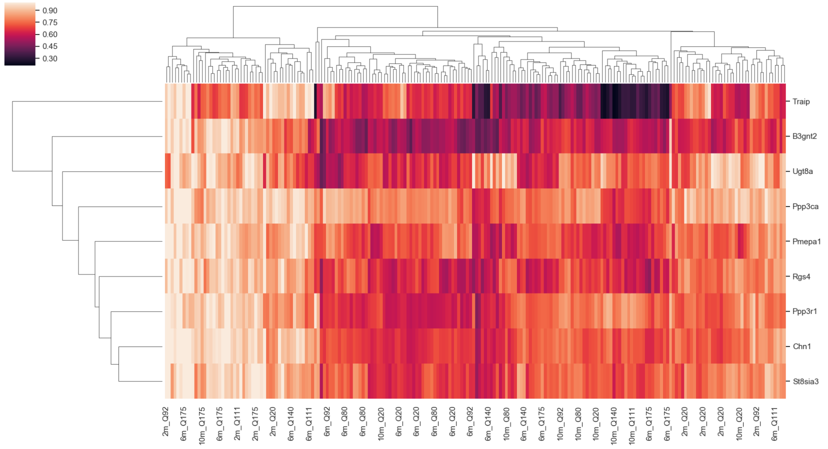


Fig 6. Clusterheatmap of the selected 9 differentially expressed disease-associated genes.

binding and calmodulin binding. A previous report uncovered Ppp3r1 related to higher cerebrospinal fluid tau level. Previous study showed single-nucleotide polymorphisms (SNPs) located in gene regulatory subunit of Ppp3r1 (rs1868402). Our study showed Ppp3r1 was also a critical differentially expressed gene, consistent with previous research.

Chn1 (chimerin 1) locate at 2q31.1. Duane Retraction Syndrome 2 and Duane Syndrome Type 2 are associated with CHN1. Gene ontology annotations related to this gene include GTPase activator and ephrin receptor binding.

St8sia3 (ST8 Alpha-N-Acetyl-Neuraminide Alpha-2,8-Sialyltransferase 3) locate at 18q21.31. Gene ontology annotations for this gene are sialyltransferase activity and alpha-N-acetylneuraminate alpha-2,8-sialyltransferase activity.

Pathomechanism analysis

Due to the clinical heterogeneity of chronic complex neurodegenerative disease, it is imperative and necessary to conduct pathway analysis to identify key pathways associated with disease.

We further annotated the 7 genes using GeneAnalytics (https://ga.genecards.org/). The 7 genes mainly expressed in globus pallidus, caudate nucleus, cerebral cortex, subthalamic nucleus, pons, HyStem+TGFbeta3+GDF5-induced SK11 cells, and medulla oblongata in brain. Besides, it has been reported that seckel syndrome 9, epileptic encephalopathy, infantile or early childhood 1, arthrogryposis, cleff palate, craniosynostosis, and impaired intellectual development, Duane retraction sybdrome 2, Duane syndrome type 2, Duane retraction syndrome, seckel syndrome, undetermined early-onset epileptic encephalopathy, autosomal dominant non-sybdromic intellectual disability are associated with these 7 genes.

Go terms for the 7 genes include calcium-dependent protein serine/threonine phosphatase activity, calcineurin complex, calcineurin-NFAT signaling cascade, cyclosporine a binding, calmodulin binding, response to amphetamine, wnt signaling pathway, and calcium modulating pathway.

Pathways associated with the 7 genes include DARPP-32 phosphorylation, G-AlphaQ signaling, NNOS signaling at neuronal synapses, Nur77 signaling in T-cell, initiation of transcription and translation elongation at the HIV-1 LTR, MAPK-Erk pathway and tacrolimus/cyclosporine pathway.

PPI network analysis

We used the STRING (https://string-db.org) to examine relationships among the 9 genes in protein-protein interaction networks, see fig. 7. There are close connections in this module. The module may play a key role in the pathology of Huntington's disease. Moreover, we investigate the functional enrichments of the 7 genes, which are shown in Table 4.

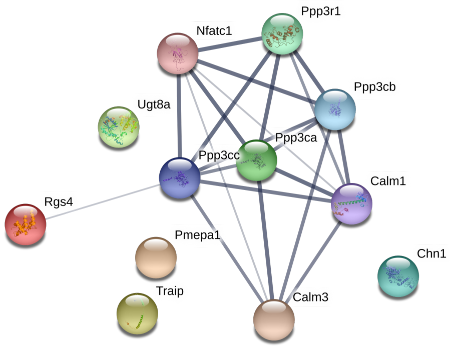


Fig 7. Protein-protein interactions among the 9 differentially expressed genes annotated by STRING.

Table 4. The functional enrichment analysis of the selected 9 differentially expressed genes.

Functional enrichments	GO-term description
	calcineurin-NFAT signaling cascade
	mitotic cell cycle phase transition
Biological Process	protein dephosphorylation
	calmodulin-dependent protein phosphatase activity
	type 3 metabotropic glutamate receptor binding
Molecular Function	phosphoprotein phosphatase activity
	calcineurin complex
	sarcolemma
Cellular Component	intracellular part
Pathways	Kaposi's sarcoma-associated herpesvirus infection
	Amphetamine addiction
	Renin secretion
	CLEC7A(Dectin-1) induces NFAT activation
UniPort keywords	protein phosphatase
	calmodulin-binding
	signal transduction inhibitor
Protein domains and features	Serine/threonine-specific protein phosphatase
	Metallo-dependent phosphatase-like
	Calcineurin-like phosphoesterase domain, ApaH type

Discussion

Benefit from discovered biomarkers, neurodegenerative disorders could be accurately diagnosed and prevented, as well as accurately treated. Nevertheless, the molecular mechanisms of many neurodegenerative disorders are still unclear. And there are no effective treatments for these disorders.

However, with the decreasing cost of sequencing technologies, large amounts of omics data become available using whole genome sequencing, RNA-seq [36], ChIP-seq [37], Methyl-seq [38], Hi-C [39,40], and ATAC-seq [41,42], etc. With all these bio-technologies, we could detect mutations in gene sequences, gene regulation and epigenetic mechanisms, reactivities of methylated and nonmethylated cytosines, chromatin interactions, and variations in gene expressions under complicated disease phenotypes. Integrative analysis of multi-omics data generated by various sequencing technologies is helpful for screening robustness biomarkers related to diseases. Diagnosis and treatment should be further personalized according to multi-dimensional information [43].

Nowadays, applications of machine learning and deep learning methods are becoming ubiquitous in the field of biomedicine and encompass not only disease gene identification, but also disease subtype classification, and transcriptional regulatory network characterization [44,45]. Besides, the accumulation of the multi-omics data and single-cell sequencing data have greatly promoted the developmental of statistical machine learning methods, evolutionaery methods, and deep learning methods. Our understanding toward the molecular mechanisms of complicated disease could be further deepen with large amounts of omics data. And the prediction accuracy could be greatly improved with the advanced algorithms.

Conclusion

In the field of biomedcine and bioinformatics, samples with labels are rare, which greatly limit the application of deep learning methods. To address this problem, we developed a generative adversarial network model. Through the game between the generator and discriminator, the prediction performance of discriminator could be greatly improved. Moreover, the variational auto-encoder can accurately capture the distribution characteristics of input data, deepen insights into laws distinguish genes of different categories.

Finally, 9 genes are selected and considered to be associated with Huntington's disease. Gene functional analysis and pathway analysis demonstrated that neurodegenerative disorders are comorbidity with many other neuropsychiatric disorders.

Supporting information

This work was supported by grants from National Natural Science Foundation of China (No. 61381671328), Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning (No. 1610000043), Innovation Research Plan supported by Shanghai Municipal Education Commission (ZXWF082101).

Acknowledgments

References

1. Mcnicholas PD, Murphy TB. Model-based clustering of microarray expression data via latent Gaussian mixture models. Bioinformatics. 2010;26(21):2705–12.

2. Krogh A, Larsson BHG, Ell S. Predicting transmembrane protein topology with a hidden Markov model:Application to complete genomes. Journal of Molecular Biology. 2001, 305(3):567-580.

3. Jiang X, Zhang H, Duan F, et al. Identify Huntington's disease associated genes based on restricted Boltzmann machine with RNA-seq data. Bmc Bioinformatics. 2017, 18(1):447.

4. Kim M, Nam J, Lee H, Ngiam J, Khosla A, et al. Multimodal deep learning. in Proc. 28th Int. Conf. Mach. Learn. 2011, 689-696.

5. Zhang Y, Lu Z. Exploring Semi-supervised Variational Autoencoders for Biomedical Relation Extraction. 2019.

6. Aalen O O. Further results on the non-parametric linear regression model in survival analysis. Statistics in Medicine. 2010, 12(17):1569-1588.

7. Altman EI, Marco G, Varetto F. Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks. Journal of Banking & Finance. 1994, 18(3):505-529.

8. Guo Y, Hastie T, Tibshirani R Regularized linear discriminant analysis and its application in microarrays. Biostatistics. 2007, 8(1):86-100.

9. Furey TS, Cristianini N, Duffy N, et al.. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics. 2000, 16(10):906-14.

10. Vaughn ML, Taylor SJ, Foy MA, et al.. Direct knowledge discovery and interpretation from a multilayer perception network which performs low-back-pain classification. Knowledge Discovery & Data Mining. 1999.

11. Liang M , Li Z , Chen T , et al. Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach. IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM, 2015.

12. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. International Conference on Neural Information Processing Systems. 2014.

13. Heusel M, Ramsauer H, Unterthiner T, et al. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. 2017.

14. Mescheder L, Geiger A, Nowozin S. Which Training Methods for GANs do actually Converge?. 2018.

15. Karras T, Laine S, Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks. 2018.

16. Roth K, Lucchi A, Nowozin S, et al. Stabilizing Training of Generative Adversarial Networks through Regularization. 2017.

17. Perarnau G, Joost V D W, Raducanu B, et al. Predicting transmembrane protein topology with a hidden Markov model:Application to complete genomes. Invertible Conditional GANs for image editing. 2016.

18. Shrivastava A, Pfister T, Tuzel O, et al. Learning from Simulated and Unsupervised Images through Adversarial Training. 2016.

19. Pan J, Ferrer CC, McGuinness K, et al. Salgan: Visual saliency prediction with generative adversarial networks. arXiv preprint arXiv:1701.01081, 2017.

20. Zhang H, Sindagi V, Patel VM. Image De-raining Using a Conditional Generative Adversarial Network. 2017.

21. Ozery-Flato M , Thodoroff P , El-Hay T. Adversarial Balancing for Causal Inference. 2018.

22. Kingma, DP, Welling, M. Auto-encoding variational bayes. In proceedings of the international conference on learning representations (ICLR).

23. Robinson, MD, Smyth, GK Moderated statistical tests for assessing differences in tag abundance. Bioinformatics. 2007, 23(21):2881—2887.

24. Robinson, MD, McCarthy, DJ, Smyth, GK. Edger: a bioconductor package for differential expression analysis of digital gene expression data.. Bioinformatics. 2010, 26(1):139–140.

25. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for rna-sequencing and microarray studies. Nucleic acids research. 2015, 43(7):47–47.

26. Hong F, Breitling R. Auto-encoding variational bayes. Bioinformatics. 2008, 24(3):374–382.

27. Wang HQ, Zheng CH, Zhao XM. Jnmfma: a joint non-negative matrix factorization meta-analysis of transcriptomics data. Bioinformatics. 2014,31(4):572–580.

28. Jiang X, Zhang H, Zhang Z, Quan X. Flexible Non-negative Matrix Factorization to Unravel Disease-related Genes. IEEE/ACM Transactions on Computational Biology Bioinformatics. 2018, PP(99):1–11.

29. Langfelder P, Cantle J P, Chatzopoulou D, et al. Integrated genomics and proteomics define huntingtin CAG length − dependent networks in mice. Nature Neuroscience. 2016, 19(4):623.

30. Yamamoto S, Jaiswal M, Charng W, et al. A drosophila genetic resource of mutants to study mechanisms underlying human genetic diseases. Cell. 2014, 159(1):200–214.

31. Ross C A, Aylward E H, Wild E J, et al. Huntington disease: natural history, biomarkers and prospects therapeutics. Nature Reviews Neurology. 2014, 10(4):204–216.

32. Kingma D P, and Welling M. Auto-encoding variational bayes. In processings of the international conference on learning representations (ICLR). 2014a.

33. kingma D, Rezende D, Mohamed S, and Welling M. Semi-supervised learning with deep generative models. In NIPS. 2014.

34. Kingma D P. Flast gradient-based inference with continuous latent variables models in auxiliary form. Technical Report. 2013.

35. Rezende D J, Mohamed S, and Wierstra D. Stochastic backpropagation and approximate inference in deep generative models. In ICML. 2014.

36. Marioni J C , Mason C E , Mane S M , et al. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Research, 2008, 18(9):1509-1517.

37. Park P J. ChIP-Seq: advantages and challenges of a maturing technology.. Nature Reviews Genetics, 2009, 10(10):669.

38. Khanna A, Czyz A, Syed F. EpiGnome Methyl-Seq Kit: a novel post-bisulfite conversion library prep method for methylation analysis.. Nature Methods, 2013, 10(10).

39. Van B N L, Erez L A, Louise W, et al. Hi-C: A Method to Study the Three-dimensional Architecture of Genomes. Journal of Visualized Experiments Jove, 2010, 39(39):292-296.

40. Belton J M, Mccord R P, Gibcus J, et al. Hi-C: A comprehensive technique to capture the conformation of genomes. Methods, 2012, 58(3):268-276.

41. Buenrostro J D , Wu B , Litzenburger U M , et al. Single-cell chromatin accessibility reveals principles of regulatory variation.. Nature, 2015, 523(7561):486-490.

42. Buenrostro J D, Wu B, Chang H Y, et al. ATAC[U+2010]seq: A Method for Assaying Chromatin Accessibility Genome[U+2010]Wide. Current Protocols in Molecular Biology, 2015, 109(1):21.29.1-21.29.9.

43. Deo R C. Machine Learning in Medicine. Circulation, 2015, 132(20):1920-1930.

44. Camacho, Diogo M. et al. Next-Generation Machine Learning for Biological Networks. Cell, 2018, 173(7): 1581–92.

45. Zou, James et al. A Primer on Deep Learning in Genomics. Nature Genetics, 2019, 51(1): 12–18.