frontiers
in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 20 July 2021
doi: 10.3389/fcell.2021.712020

Identification of Novel Susceptible Genes of Gastric Cancer Based on Integrated Omics Data

Huang Yaoxing, Yu Danchun, Sun Xiaojuan, Jiang Shuman, Yan Qingqing and Jia Lin**

Department of Gastroenterology, Guangzhou First People's Hospital, School of Medicine, South China University of Technology, Guangzhou, China

Gastric cancer (GC) is one of the most common causes of cancer-related deaths in the world. This cancer has been regarded as a biological and genetically heterogeneous disease with a poorly understood carcinogenesis at the molecular level. Thousands of biomarkers and susceptible loci have been explored via experimental and computational methods, but their effects on disease outcome are still unknown. Genome-wide association studies (GWAS) have identified multiple susceptible loci for GC, but due to the linkage disequilibrium (LD), single-nucleotide polymorphisms (SNPs) may fall within the non-coding region and exert their biological function by modulating the gene expression level. In this study, we collected 1,091 cases and 410,350 controls from the GWAS catalog database. Integrating with gene expression level data obtained from stomach tissue, we conducted a machine learning-based method to predict GC-susceptible genes. As a result, we identified 787 novel susceptible genes related to GC, which will provide new insight into the genetic and biological basis for the mechanism and pathology of GC development.

Keywords: gastric cancer, GWAS, integrated omics data, a machine learning based method, biomarkers

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Shihua Zhang,
Wuhan University of Science and Technology, China
Haixiu Yang,
Harbin Medical University, China

****Correspondence:***
Huang Yaoxing
huangyaoxing@sina.com
Jia Lin
sqt661841@163.com

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *19 May 2021*
Accepted: *23 June 2021*
Published: *20 July 2021*

Citation:
Yaoxing H, Danchun Y, Xiaojuan S, Shuman J, Qingqing Y and Lin J (2021) Identification of Novel Susceptible Genes of Gastric Cancer Based on Integrated Omics Data. Front. Cell Dev. Biol. 9:712020. doi: 10.3389/fcell.2021.712020

INTRODUCTION

Gastric cancer (GC) is one of the most common malignant dangerous neoplasms and fatal diseases in the world. It has been reported that there were approximately one million newly diagnosed gastric carcinoma cases, which caused 780,000 deaths in 2018 (Bray et al., 2018). It is notable that nearly half of the GC incidences occurred in the Asian region, which partly resulted from the diverse hereditary background, behavioral factors, and the spread of and infection by *Helicobacter pylori* (Chen et al., 2016; Zheng et al., 2017). A lot of work has been done to improve the diagnosis and therapy of GC. However, the survival rate of GC patients remains poor at approximately 30% in the recent 5 years (DeSantis et al., 2014). Therefore, many efforts have been made to discover new biomarkers to help in staging and in prognosis of the tumor diagnosis, which could help in improving early diagnosis and prognostic prediction of GC (Ludwig and Weinstein, 2005).

Disease gene prediction is a task of identifying the significant susceptible genes related to diseases. There have been a variety of approaches proposed, such as annotation-based approaches, network-based approaches, and machine learning-based approaches. Annotation-based approaches, such as prioritization of candidate genes using statistics (POCUS) (Turner et al., 2003), SUSPECTS (Adie et al., 2006), Endeavor (Aerts et al., 2006), and Transcriptome Ontology Pathway PubMed based prioritization of Genes (ToppGene) (Chen et al., 2007), are proposed based on annotating the genes with respect to biological structures or functions then comparing the annotations with known disease causal genes. However, these methods have a limitation of failing to capture the indirect relationships between the genes that may have common features or functions but are still not annotated. Network-based methods are proposed to overcome this by utilizing the large scale of interactome data between cellular molecules covering most of the genome and proteome (Wang et al., 2011).

Machine learning techniques have been applied to solve various biomedical problems (Lu and Zhao, 2019; Zhao et al., 2020a,b), such as pattern recognition (Barral et al., 2012; Cui et al., 2020), classification (Basford et al., 2013; Zhao et al., 2021a), prediction of drug target (Ding et al., 2014; Tianyi et al., 2020), and genome annotation (Yip et al., 2013). Thus, there is no doubt that machine learning methods have been applied in disease-associated genes prediction (Calvo et al., 2006; Xu and Li, 2006). In recent years, emerging evidences have illustrated the essential role of single-nucleotide polymorphisms (SNPs) in GC development and progression. Since genome-wide association studies (GWAS) are a widely known powerful approach to explore complex susceptible variants of diseases, many studies have reported a number of susceptibility loci associated with GC through GWAS analysis; however, they can only explain a small fraction of GC heritability (Wang et al., 2017; Park et al., 2019). Moreover, most disease-related SNPs identified by GWAS fall into intergenic or non-coding regions, which may influence the process of pathogenesis by modulating the expression level of target genes (Maurano et al., 2012). However, genetic variants are still powerful and high-quality biomarkers for screening GC susceptibility (Mocellin et al., 2015).

In addition, it has been proven that gene expression is significantly related to diseases (Zhao et al., 2021b). Expression quantitative trait locus (eQTL) analysis has been regarded as a powerful approach to provide prior weights for the statistical analysis of new causal SNP identification and prioritize SNPs or genes for further validation (Li et al., 2013b). Due to the theory of linkage disequilibrium (LD), which is reflected by the non-random association of alleles of different loci, it can be inferred that SNPs can regulate the pathologies of diseases by modulating the expression level of target genes. However, most studies select the representative SNPs by their closest located gene, which may inevitably obscure the genetic effect between that candidate gene and the trait. Thus, integrating GWAS data and eQTL data can help us to detect the genetic mechanism of complex disease. The Genotype-Tissue Expression (GTEx) project has provided the largest comprehensive public database of whole-genome and transcriptome sequencing data to help better understand the effects and molecular mechanism of function variations.

Considering the fact that regulatory causal SNPs may exert their function by affecting their target gene expression, we collected 1,091 GC cases and 410,350 controls from GWAS catalog database and eQTL summary data from stomach tissues in the GTEx database (Rashkin et al., 2020). We then extracted gene features from both GWAS summary data and eQTL data and integrated them as a 10D vector to represent gene feature, then we performed several machine learning methods to assess the classification performance of the models and selected random forest (RF) classifier based on its excellent performance. We identified 787 novel susceptible genes related to GC which may help provide new insights into the mechanism of GC.

MATERIALS AND METHODS

GC GWAS Datasets

In this study, we obtained a GC GWAS dataset consisting of 1,091 cases and 410,350 controls from the UK Biobank (Rashkin et al., 2020). All the subjects were genotyped with Affymetrix Genome-Wide Human SNP Array. Eleven significant susceptible SNPs were identified related to GC and esophageal carcinoma with p-value $< 1 \times 10^{-6}$. However, they identified 9,986,610 susceptible loci associated with stomach and esophageal cancer in total, which is utilized in our study to be further identified.

Tissue eQTL Dataset

Expression level-associated SNPs in stomach tissues were obtained from the GTEx v8 database. Genotyping was performed utilizing Illumina HumanOmni 5 and 2.5 M. And transcriptome dataset was generated by using Affymetrix Expression Array or Illumina TruSeq RNA sequencing. As a result, 24,291 susceptible loci were identified based on gene expression level.

GC-Related Genes and Candidate Genes

After obtaining both omics summary data, we obtained 5,632 gastric-related genes from DisGeNET database, which are considered as positive genes for machine learning methods (Bauer-Mehren et al., 2010). Then we downloaded the human gene networks from HumanNet v2.0, which was a database of human gene networks illustrating gene–gene interactions (Hwang et al., 2018). After filtering the genes related to the GC causal genes obtained from DisGeNET, we obtained 3,227 genes whose correlation scores were <1, which indicate that they are negative genes, and the rest of the genes are regarded candidate genes which may have association with GC.

Feature Extraction

We first obtained GWAS dataset and eQTL dataset associated with GC from the GWAS catalog database and the GTEx database, which represent the gene feature from the aspect of phenotype and transcription, respectively. Since the LD correction structure means that the majority of the identified variants associated with the traits frequently point to the regions where many genes are located, it is extremely difficult to prioritize among these susceptible genes to identify the most functionally relevant causal genes merely based on GWAS data. It is widely known that SNPs can exert their regulation function by modulating the expression level of target genes, which may further have a significant influence on the phenotype. Many studies have applied analytical approaches to integrate eQTL and GWAS data to detect the causal genes associations with complex traits (Giambartolomei et al., 2014; Gusev et al., 2016). However, to our knowledge, there is no method utilizing machine learning classifiers to prioritize susceptible genes of GC based on integrated GWAS and eQTL summary data.

After obtaining both omics summary data, we obtained gastric-related genes from DisGeNET database (Piñero et al., 2016). We first performed a data preprocessing process to transform the names of genes downloaded from different databases. We used an R package to obtain the detailed information of the genes, such as chromosome, start position, and end position information of these genes. After mapping these genes to the GWAS and eQTL data based on the location information, we finally got 5,633 genes regarded as a training set. According to the mapping information, we kept the genes with at least one susceptible loci identified by GWAS analysis. After prioritizing the SNPs by their *p*-value obtained by GWAS analysis, we used the *p*-value of the top five SNPs related to each gene as a 5D phenotype-based feature vector. For those genes with less than five associated SNPs, the feature vector is filtered with 1, which means the gene has no correlation with GC. Thus, the phenotype-based feature vector of G_i can be denoted as follows:

$$G_i^P = \left[P_p^1,\ P_p^2,\ P_p^3,\ \ P_p^4,\ P_p^5 \right] \tag{1}$$

After obtaining the top five SNPs associated with these genes, these SNPs can be mapped to the stomach tissue eQTL data. Based on the same method, we use the *p*-value of those SNPs successfully mapped in eQTL data to represent a 5D transcriptome-based feature vector of each gene. For those SNPs not mapped, we all use 1 to fill up the feature vector. Thus, the transcriptome-based feature vector can be denoted as follows:

$$G_i^T = \left[P_T^1,\ P_T^2,\ P_T^3,\ \ P_T^4,\ P_T^5 \right] \tag{2}$$

Thus, each gene feature can be represented as a 10D feature vector based on the integrated omics data.

Gene Prediction Using Random Forest Classifier

To date, the binary classification methods have been widely applied in disease causal gene prediction problems, such as naïve Bayesian classifier (NB), support vector machine (SVM), RF, and some deep learning methods such as convolutional neural network (CNN), graph neural network (GCN), and deep neural network (DNN). In this work, we used a RF classifier to predict GC causal genes. In order to evaluate the RF classifier, we performed a 10-fold cross-validation on the training set. The main workflow is shown in **Figure 1**. First, the 5,633 genes are randomly divided into 10 groups; 9 of them were chosen to be training samples, and the last one is the testing set for a total of 10 training iterations. Grid searches were performed to obtain the best performance of the parameters of the RF classifier. The final statistical results are averaged after 10 iterations. The receiver operating characteristic (ROC) curve and the area under the curve (AUC) are utilized to assess the performance of the classifier. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings to illustrate the diagnostic ability of a binary classifier system. This measure is related to the evaluation criteria TPR, which can be denoted as follows:

$$\text{TPR} = \frac{TP}{TP + FN} \tag{3}$$

where *TP* means true positive conditions, and *FN* means the false negative conditions. TPR is also known as sensitivity or recall. Another measurement is true negative rate (TNR), which is also known as specificity or selectivity; TNR can be denoted as follows:

$$\text{TNR} = \frac{TN}{TN + FP} \tag{4}$$

Another measurement is FPR, which can be denoted as follows:

$$\text{FPR} = \frac{FP}{FP + TN} = 1\text{−TNR} \tag{5}$$

where *FP* indicates the false positive results and *TN* indicates the true negative results. AUC means the AUC. The AUC of each iteration is shown in **Table 1**. After the model is trained, we use the model to predict the candidate genes.

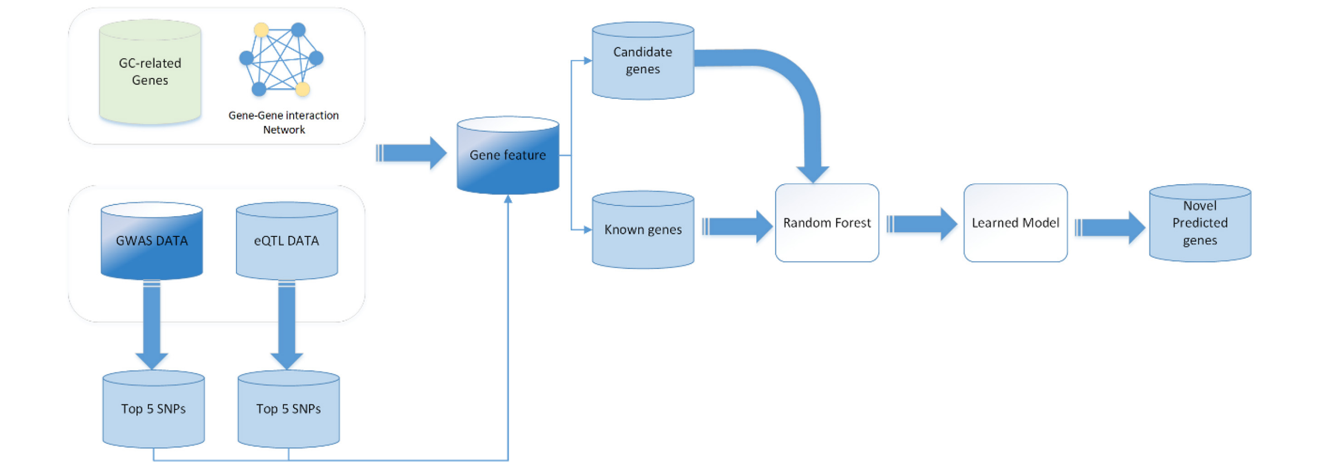

FIGURE 1 | Workflow of gastric cancer-related genes prediction.

TABLE 1 | Area under the curve of 10-fold cross-validation.

	AUC
1	0.881
2	0.91
3	0.894
4	0.886
5	0.878
6	0.872
7	0.896
8	0.892
9	0.902
10	0.909
Average	0.892

RESULTS

Performance Comparison Over All Known Disease Genes

In this study, we use all known GC-related genes obtained from DisGeNET as positive training samples, and genes with correlation scores under 1 were obtained from HumanNet v2.0 as negative training samples. We compared the predictive performance of RF, SVM, NB, and DNN. After 10-fold cross-validation of each method, the ROC curve and average AUC value is shown in **Figure 2**. The AUC of RF is 0.892, followed by the AUC of 0.811 of DNN, an AUC of 0.753 of SVM, and an AUC of 0.593 of NB, which means RF did the best performance in disease gene classification. For details, the AUC value of each iteration in RF classifier is shown in **Table 1**.

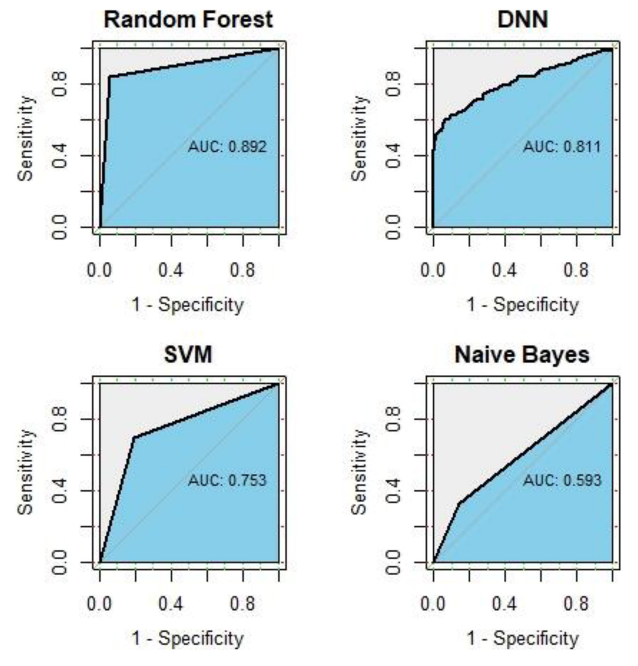

FIGURE 2 | Comparison of prediction performance.

TABLE 2 | Predicted susceptible genes by random forest.

Symbol	Ensembl ID	Symbol	Ensembl ID
ANLN	ENSG00000011426	CD38	ENSG00000004468
TYROBP	ENSG00000011600	PDK4	ENSG00000004799
MATR3	ENSG00000015479	RARB	ENSG00000077092
NCDN	ENSG00000020129	ITGA2B	ENSG00000005961
TYMP	ENSG00000025708	CRLF1	ENSG00000006016
ALG1	ENSG00000033011	NRAS	ENSG00000213281
KRAS	ENSG00000133703	GTF2IRD1	ENSG00000006704
CDH1	ENSG00000039068	CACNA2D2	ENSG00000007402
BEST2	ENSG00000039987	DNAJC11	ENSG00000007923
TNFRSF17	ENSG00000048462	RPS20	ENSG00000008988
ADAMTS6	ENSG00000049192	GIPR	ENSG00000010310
PLAUR	ENSG00000011422	SLC6A7	ENSG00000011083

Case Study

We obtained 7,406 candidate genes from HumanNet v2.0, and then extracted the feature representation of each gene. Then we used the trained RF classifier to predict the prioritizing genes associated with GC. The top 24 susceptible genes are shown in **Table 2**. From **Table 2**, 11 of the 24 predicted genes are reported to have direct or indirect association with GC. For example, CD38 has been determined to be expressed at higher levels in the IL-10-producing Breg cells of GC patients (Wang et al., 2015). The study of Cheng et al. found that compared to control tissues, RARB messenger RNAs were significantly reduced in human gastric tumor samples (Cheng et al., 2013). As an indirect evidence, Wen et al. found that NRAS can be a target gene of miR-26a to improve the sensitivity of GC cells to cisplatin-based chemotherapies, which can be an evidence for the potential function of NRAS in chemotherapy for GC (Wen et al., 2015). GIPR, short for gastric inhibitory polypeptide receptor, has been regarded as a promising target for imaging and therapy in gastric and neuroendocrine tumors, and it has also been reported that GIPR is significantly overexpressed in stomach tissue compared with normal tissue (Sherman et al., 2013, 2014). In recent years, many studies have shown that PLAUR can be an effective prognostic biomarker and potential therapeutic target for GC due to the fact that the suppression of PLAUR could sensitize cancer cell death by inducing DNA damages (Li et al., 2013a; Ai et al., 2020). ANLN is a conserved actin-binding protein that exerts its functions in cytoskeletal dynamics during cell division and may affect cancer progression through Wnt/B-catenin pathway in GC.

CONCLUSION

Gastric cancer is one of the most malignant neoplasms in human health around the world causing approximately 10% of all cancer deaths. The main therapeutic strategies of GC include two ways: surgery and chemotherapeutic regimens. Thus, it is important to identify the susceptible genes in order to better understand the pathologies of the disease, which can further help in drug designing. Machine learning methods have been used in predicting the functions of unclassified or unannotated genes by utilizing genomic features (Schläpfer et al., 2017). In this study, we combined machine learning methods with genetic association data (GWAS analysis) and gene expression data (eQTL) to advance our understanding of GC etiology and pathology. Considering the LD, genes located close to susceptible loci identified by GWAS analysis may not be the causal genes of the disease. Since SNPs may also influence the expression level of the gene, the genes with different genotypes of the genetic variant will show differences in phenotype, which means that SNPs can also show effects on the diseases or traits. Therefore, we performed a RF classier on the collected 1,091 cases and 410,350 controls from GWAS dataset, integrated with stomach tissue eQTL data, to identify genes whose expression levels were associated with GC due to its causality. Compared to three other widely used binary classifiers, SVM, NB, and DNN, RF has the best performance in classifying GC-related genes. Since the overall importance score from the RF classifier is a sum of multiple individual importance scores, and each individual importance score is obtained from an average over multiple trees and cross-validations, the gene importance score can be regarded as not being affected by LD.

It is widely known that the accuracy of the prediction of genetic risk of complex diseases varies greatly between different diseases due to the heritability, phenotype, and the power and the amount of reported variants. Though GC is the most common cancer and causes a high mortality rate around the world, most studies only focused on the prognosis and treatment of GC. In this study, we identified 787 novel susceptible genes related to GC and focused on the top 24 of the susceptible genes. We found that CD38, RARB, NRAS, GIPR, PLAUR, ANLN, etc. have strong association with GC and have been reported to be related with GC indirectly; for example, they impact other pathways or exert their function by cooperating with other genes. However, we identified 787 novel susceptible genes related to GC, which is helpful in further studies to understand the etiology and pathology of GC and may be a strong theoretical basis for drug designing.

We used a machine learning method to identify GC-related genes. Although we have proven the effectiveness of our method by cross-validation and accuracy of our results by case study, biological experiments are still necessary to further reveal pathogenic mechanisms.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. This data can be found here: GWAS dataset: downloaded from GWAS Catalog, accession: GCST90011807.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements. Written informed consent was not obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

HY and JL designed the study. HY, YD, and SX interpreted the data. HY, JS, and YQ analyzed the results. HY, YD, and JL were major contributors in writing and revising the manuscript. All authors read and approved the final manuscript.

REFERENCES

Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J., and Pickard, B. S. (2006). Suspects: enabling fast and effective prioritization of positional candidates. *Bioinformatics* 22, 773–774. doi: 10.1093/bioinformatics/btk031

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., et al. (2006). Gene prioritization through genomic data fusion. *Nat. Biotechnol.* 24, 537–544. doi: 10.1038/nbt1203

Ai, C., Zhang, J., Lian, S., Ma, J., Győrffy, B., Qian, Z., et al. (2020). FOXM1 functions collaboratively with PLAU to promote gastric cancer progression. *J. Cancer* 11, 788–794. doi: 10.7150/jca.37323

Barral, S., Bird, T., Goate, A., Farlow, M., Diaz-Arrastia, R., Bennett, D., et al. (2012). Genotype patterns at PICALM, CR1, BIN1, CLU, and APOE genes are associated with episodic memory. *Neurology* 78, 1464–1471. doi: 10.1212/wnl.0b013e3182553c48

Basford, K. E., McLachlan, G. J., and Rathnayake, S. I. (2013). On the classification of microarray gene-expression data. *Brief. Bioinform.* 14, 402–410. doi: 10.1093/bib/bbs056

Bauer-Mehren, A., Rautschka, M., Sanz, F., and Furlong, L. I. (2010). DisGeNET: a cytoscape plugin to visualize, integrate, search and analyze gene–disease networks. *Bioinformatics* 26, 2924–2926. doi: 10.1093/bioinformatics/btq538

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Calvo, S., Jain, M., Xie, X., Sheth, S. A., Chang, B., Goldberger, O. A., et al. (2006). Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat. Genet.* 38, 576–582. doi: 10.1038/ng1776

Chen, J., Xu, H., Aronow, B. J., and Jegga, A. G. (2007). Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics* 8:392. doi: 10.1186/1471-2105-8-392

Chen, W., Zheng, R., Baade, P. D., Zhang, S., Zeng, H., Bray, F., et al. (2016). Cancer statistics in China, 2015. *CA Cancer J. Clin.* 66, 115–132. doi: 10.3322/caac.21338

Cheng, A. S., Li, M. S., Kang, W., Cheng, V. Y., Chou, J. L., Lau, S. S., et al. (2013). *Helicobacter* pylori causes epigenetic dysregulation of FOXD3 to promote gastric carcinogenesis. *Gastroenterology* 144, 122–133. doi: 10.1053/j.gastro.2012.10.002

Cui, S. J., Ganjawala, T. H., Abrams, G. W., and Pan, Z. H. (2020). Effect of proteasome inhibitors on the AAV-mediated transduction efficiency in retinal bipolar cells. *Curr. Gene Ther.* 19, 404–412. doi: 10.2174/1566523220666200211111326

DeSantis, C. E., Lin, C. C., Mariotto, A. B., Siegel, R. L., Stein, K. D., Kramer, J. L., et al. (2014). Cancer treatment and survivorship statistics, 2014. *CA Cancer J. Clin.* 64, 252–271. doi: 10.3322/caac.21235

Ding, H., Takigawa, I., Mamitsuka, H., and Zhu, S. (2014). Similarity-based machine learning methods for predicting drug–target interactions: a brief review. *Brief. Bioinform.* 15, 734–747. doi: 10.1093/bib/bbt056

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., et al. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* 10:e1004383. doi: 10.1371/journal.pgen.1004383

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* 48, 245–252. doi: 10.1038/ng.3506

Hwang, S., Kim, C. Y., Yang, S., Kim, E., Hart, T., Marcotte, E. M., et al. (2018). HumanNet v2: human gene networks for disease research. *Nucleic Acids Res.* 47, D573–D580. doi: 10.1093/nar/gky1126

Li, D., Wei, P., Peng, Z., Huang, C., Tang, H., Jia, Z., et al. (2013a). The critical role of dysregulated FOXM1–PLAUR signaling in human colon cancer progression and metastasis. *Clin. Cancer Res.* 19, 62–72. doi: 10.1158/1078-0432.ccr-12-1588

Li, L., Kabesch, M., Bouzigon, E., Demenais, F., Farrall, M., Moffatt, M. F., et al. (2013b). Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma. *Front. Genet.* 4:103. doi: 10.3389/fgene.2013.00103

Lu, X. X., and Zhao, S. Z. (2019). Gene-based therapeutic tools in the treatment of cornea disease. *Curr. Gene Ther.* 19, 7–19. doi: 10.2174/1566523219666181213120634

Ludwig, J. A., and Weinstein, J. N. (2005). Biomarkers in cancer staging, prognosis and treatment selection. *Nat. Rev. Cancer* 5, 845–856. doi: 10.1038/nrc1739

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 1190–1195. doi: 10.1126/science.1222794

Mocellin, S., Verdi, D., Pooley, K. A., and Nitti, D. (2015). Genetic variation and gastric cancer risk: a field synopsis and meta-analysis. *Gut* 64, 1209–1219. doi: 10.1136/gutjnl-2015-309168

Park, B., Yang, S., Lee, J., Woo, H. D., Choi, I. J., Kim, Y. W., et al. (2019). Genome-wide association of genetic variation in the PSCA gene with gastric cancer susceptibility in a Korean population. *Cancer Res. Treat.* 51, 748–757. doi: 10.4143/crt.2018.162

Piñero, J., Bravo, À, Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., et al. (2016). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* 45, D833–D839. doi: 10.1093/nar/gkw943

Rashkin, S. R., Graff, R. E., Kachuri, L., Thai, K. K., Alexeeff, S. E., Blatchins, M. A., et al. (2020). Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. *Nat. Commun.* 11:4423. doi: 10.1038/s41467-020-18246-6

Schläpfer, P., Zhang, P., Wang, C., Kim, T., Banf, M., Chae, L., et al. (2017). Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiol.* 173, 2041–2059. doi: 10.1104/pp.16.01942

Sherman, S. K., Carr, J. C., Wang, D., O'Dorisio, M. S., O'Dorisio, T. M., and Howe, J. R. (2013). Gastric inhibitory polypeptide receptor (GIPR) is a promising target for imaging and therapy in neuroendocrine tumors. *Surgery* 154, 1206–1214. doi: 10.1016/j.surg.2013.04.052

Sherman, S. K., Maxwell, J. E., Carr, J. C., Wang, D., O'Dorisio, M. S., O'Dorisio, T. M., et al. (2014). GIPR expression in gastric and duodenal neuroendocrine tumors. *J. Surg. Res.* 190, 587–593. doi: 10.1016/j.jss.2014.01.044

Tianyi, Z., Yang, H., Valsdottir, L. R., Tianyi, Z., and Jiajie, P. (2020). Identifying drug–target interactions based on graph convolutional network and deep neural network. *Brief. Bioinform.* 22, 2141–2150. doi: 10.1093/bib/bbaa044

Turner, F. S., Clutterbuck, D. R., and Semple, C. A. (2003). POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol.* 4:R75. doi: 10.1186/gb-2003-4-11-r75

Wang, W., Yuan, X., Chen, H., Xie, G., Ma, Y., Zheng, Y., et al. (2015). CD19+ CD24hiCD38hiBregs involved in downregulate helper T cells and upregulate regulatory T cells in gastric cancer. *Oncotarget* 6, 33486–33499. doi: 10.18632/oncotarget.5588

Wang, X., Gulbahce, N., and Yu, H. (2011). Network-based methods for human disease gene prediction. *Brief. Funct. Genomics* 10, 280–293. doi: 10.1093/bfgp/elr024

Wang, Z., Dai, J., Hu, N., Miao, X., Abnet, C. C., Yang, M., et al. (2017). Identification of new susceptibility loci for gastric non-cardia adenocarcinoma: pooled results from two Chinese genome-wide association studies. *Gut* 66, 581–587. doi: 10.1136/gutjnl-2015-310612

Wen, L., Cheng, F., Zhou, Y., and Yin, C. (2015). MiR-26a enhances the sensitivity of gastric cancer cells to cisplatin by targeting NRAS and E2F2. *Saudi J. Gastroenterol.* 21, 313–319. doi: 10.4103/1319-3767.166206

Xu, J., and Li, Y. (2006). Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics* 22, 2800–2805. doi: 10.1093/bioinformatics/btl467

Yip, K. Y., Cheng, C., and Gerstein, M. (2013). Machine learning and genome annotation: a match meant to be? *Genome Biol.* 14:205. doi: 10.1186/gb-2013-14-5-205

Zhao, T., Hu, Y., and Cheng, L. (2020a). Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep learning approaches. *Brief. Bioinform.* 13:bbaa212. doi: 10.1093/bib/bbaa212

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020b). DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. *Bioinformatics* 36, 4466–4472. doi: 10.1093/bioinformatics/btaa428

Zhao, T., Liu, J., Zeng, X., Wang, W., Li, S., Zang, T., et al. (2021a). Prediction and collection of protein–metabolite interactions. *Brief. Bioinform.* 12:bbab014. doi: 10.1093/bib/bbab014

Zhao, T., Lyu, S., Lu, G., Juan, L., Zeng, X., Wei, Z., et al. (2021b). SC2disease: a manually curated database of single-cell transcriptome for human diseases. *Nucleic Acids Res.* 49, D1413–D1419. doi: 10.1093/nar/gkaa838

Zheng, R., Zeng, H., Zhang, S., and Chen, W. (2017). Estimates of cancer incidence and mortality in China, 2013. *Chin. J. Cancer* 36:66. doi: 10.1186/s40880-017-0234-3

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.