frontiers in Genetics

METHODS
published: 05 February 2020
doi: 10.3389/fgene.2019.01401

Evaluation of Pathway Activation for a Single Sample Toward Inflammatory Bowel Disease Classification

Xingyi Li[1], Min Li[1], Ruiqing Zheng[1], Xiang Chen[1], Ju Xiang[1,2], Fang-Xiang Wu[3] and Jianxin Wang[1]*

[1] School of Computer Science and Engineering, Central South University, Changsha, China, [2] Neuroscience Research Center & Department of Basic Medical Sciences, Changsha Medical University, Changsha, China, [3] Department of Mechanical Engineering and Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, SK, Canada

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Meng Zhou,
Wenzhou Medical University, China
Xiangrong Liu,
Xiamen University, China
Zhi-Ping Liu,
Shandong University, China

***Correspondence:**
Min Li
limin@mail.csu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *29 September 2019*
Accepted: *23 December 2019*
Published: *05 February 2020*

Citation:
Li X, Li M, Zheng R, Chen X, Xiang J, Wu F-X and Wang J (2020) Evaluation of Pathway Activation for a Single Sample Toward Inflammatory Bowel Disease Classification. Front. Genet. 10:1401. doi: 10.3389/fgene.2019.01401

Since similar complex diseases are much alike in clinical symptoms, patients are easily misdiagnosed and mistreated. It is crucial to accurately predict the disease status and identify markers with high sensitivity and specificity for classifying similar complex diseases. Many approaches incorporating network information have been put forward to predict outcomes, but they are not robust because of their low reproducibility. Several pathway-based methods are robust and functionally interpretable. However, few methods characterize the disease-specific states of single samples from the perspective of pathways. In this study, we propose a novel framework, Pathway Activation for Single Sample (PASS), which utilizes the pathway information in a single sample way to better recognize the differences between two similar complex diseases. PASS can mainly be divided into two parts: for each pathway, the extent of perturbation of edges and the statistic difference of genes caused by a single disease sample are quantified; then, a novel method, named as an AUCpath, is applied to evaluate the pathway activation for single samples from the perspective of genes and their interactions. We have applied PASS to two main types of inflammatory bowel disease (IBD) and widely verified the characteristics of PASS. For a new patient, PASS features can be used as the indicators or potential pathway biomarkers to precisely diagnose complex diseases, discover significant features with interpretability and explore changes in the biological mechanisms of diseases.

Keywords: similar complex diseases, pathway activation, single sample, inflammatory bowel disease, pathway biomarkers

INTRODUCTION

Complex diseases threaten human health and life quality. Similar complex diseases make the early diagnosis of patients more difficult due to similar clinical symptoms. Therefore, mining effective biological information to accurately discriminate between similar complex diseases has become the most important research area of biomedicine. In the previous research, several methods based on a

single biological network, such as the metabolic network, regulatory network, or protein–protein interaction (PPI) network, have been put forward to aid in disease prediction, diagnosis, prognosis, and so on (Winter et al., 2012; Cun and Fröhlich, 2013). Nevertheless, these methods are not robust because of the low reproducibility (Yousefi and Dougherty, 2012; Amar et al., 2015; Choi et al., 2017) that results from the cellular heterogeneity within tissues, the heterogeneity of samples, and errors of measuring technologies.

Since genes generally take effect synergistically by forming functional modules, inferring features related to disease classification at the functional level can effectively ameliorate the adverse effects of heterogeneity and obtain more reproducible markers. Some methods utilize Gene Ontology (Ashburner et al., 2000) to differentiate disease states (Zhang et al., 2017) while others integrate pathway information. Pathways reflect biological processes within cells, such as metabolism, signaling, and growth cycles, and markers identified based on pathway information can thus maintain functional interpretability (Haider et al., 2018). Moreover, the occurrence and progression of complex diseases, such as inflammatory bowel disease (IBD), are often related to the dysregulation of significant pathways. Discovering the involved pathways and quantifying their disorders are of great significance in understanding complex diseases (Bild et al., 2006; Thomas et al., 2008; Markert et al., 2011; Drier et al., 2013).

A series of methods for disease classification integrate pathway information from the Molecular Signatures Database (MSigDB) (Subramanian et al., 2005) or Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000). Several works extract significant features from the genes along pathways to distinguish diseases (Huang et al., 2003; Bild et al., 2006; Lee et al., 2008a; Young and Craft, 2016). Although these works can combine pathway information to classify diseases effectively, they only regard a pathway as a set of genes and ignore the edge information between genes, which may lead to the loss of important information related to diseases. To overcome this problem, some methods for analyzing the intrinsic structures of pathways and integrating topological characteristics of pathways have been proposed (Liu et al., 2013; Han et al., 2017). These existing algorithms can effectively utilize the topological information of pathways to predict disease status. Nevertheless, none of them assesses condition-specific states for each patient from a pathway perspective, but this is essential to revealing the molecular mechanisms of complex diseases at the system level.

By analyzing the high-dimensional information of expression data and the differential distribution (i.e., volcano distribution) of a single patient against a given number of normal samples (Liu et al., 2016), we propose a novel framework to classify two similar complex diseases by evaluating the pathway activation based on single sample analysis. Our method consists of two steps: (1) a fully connected network for each pathway is constructed and the perturbation of each edge in the network caused by the introduction of each disease sample is evaluated. For all genes in the pathways, the statistical difference of gene expression between a single disease sample and normal samples is evaluated; (2) a novel method named as AUCpath is introduced to evaluate the pathway activation for single sample (PASS) of each pathway from both node and edge aspects, which converts the high-dimensional, small-sample gene expression matrix into a PASS matrix. Finally, a random forest classifier based on PASS features is built to examine the classification performance.

We applied PASS to classify ulcerative colitis (UC) and Crohn's disease (CD) (Ananthakrishnan, 2015). UC and CD have many common clinical features, such as abdominal pain, diarrhea, recurrent episodes, and so on. They are therefore collectively referred to as IBD. IBD is a special kind of intestinal inflammatory disease caused by common factors such as genetics, environmental triggers, immunoregulatory defects, and microbial exposure (Hanauer, 2006). Currently, there is no gold standard for discriminating UC and CD, but the responses and effects after medication of these two complex diseases are not the same (Akobeng et al., 2016; Baumgart and Sandborn, 2007), and this has motivated many attempts to understand the differences in the molecular characteristics between these two similar complex diseases at the tissue level (Lawrance et al., 2001; Burczynski et al., 2006; Wu et al., 2007). The improved understanding of the differential mechanisms of UC and CD from a molecular perspective can improve the diagnostic accuracy and have the potential to improve the therapeutic effect and the success rate of clinical trials.

We compare our method with seven network-based, GO-based, and pathway-based methods, respectively, and obtain prominent performance against these methods. In addition, our experimental results showed that our method can elucidate the molecular mechanism of UC and CD and has the potential to identify biomarkers with functional interpretability.

MATERIALS AND METHODS

Dataset and Preprocessing

We downloaded two pediatric datasets and three adult datasets from the Gene Expression Omnibus (GEO) (Edgar et al., 2002), namely GSE9686 (Carey et al., 2007), GSE3365 (Burczynski et al., 2006), GSE36807 (Montero-Meléndez et al., 2013), GSE71730 (Gurram et al., 2016), and GSE16879 (Arijs et al., 2009). All of them contain UC, DC, and normal samples.

In order to maintain the consistency of data and reduce the impact of noise, we selected data from the same anatomical site and patients under the same conditions. We excluded samples of CD patients during treatment for GSE9686 and samples of Crohn's ileitis for GSE16879. We mapped probes to gene ID using files provided by the corresponding platforms, discarded probes corresponding to multiple genes, and chose the median when multiple probes were mapped to the same gene to eliminate the influence of measurement errors. Only genes detected in all datasets can be used for the downstream analysis. As a result, there were 11242 genes included in all five datasets. **Table 1** summarizes the above datasets.

TABLE 1 | Summary of the gene expression datasets.

Name	Healthy	UC	CD	Total genes	Type of samples	Reference	URL
GSE9686	8	5	11	15747	Pediatric samples	(Carey et al., 2007)	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE9686
GSE3365	42	26	59	12432	Adult samples	(Burczynski et al., 2006)	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3365
GSE36807	7	15	13	20486	Adult samples	(Montero-Meléndez et al., 2013)	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36807
GSE71730	10	15	22	20486	Pediatric samples	(Gurram et al., 2016)	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE71730
GSE16879	6	24	19	20486	Adult samples	(Arijs et al., 2009)	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16879

From the KEGG database, all human pathways were downloaded using the KEGGgraph package (Zhang and Wiemann, 2009). A total of 294 pathways were extracted. Each pathway consisted of a set of genes and their interactions; genes were represented by nodes, and interactions were edges in the KEGG human pathways. Genes that were not present in the expression profiles and their corresponding interactions were discarded. Considering the following analysis, pathways containing only one edge were not included. Finally, 291 pathways were retained, and these contained 3926 genes in total.

Pathway Activation for Single Sample

Pathway-based features are more robust while maintaining biological interpretability and tend to be small in number, which can prevent overfitting. In this study, we introduced a new method, called PASS, to evaluate the state of each known pathway. PASS defined the state of a pathway from the aspect of genes and regulatory links. Although it was difficult to analyze the regulatory links in the pathway for each patient, the sample-specific network (SSN) analysis provided a feasible and effective way to mine the different regulatory patterns for each patient.

In this study, we first constructed a fully connected network for each pathway. For each dataset, we analyzed the condition-specific state for each disease sample based on the pathway and thus assessed the PASS features. The schematic diagram of our framework is shown in **Figure 1**.

Statistical Difference of Edges Between Single Disease Sample and Normal Samples

For each fully connected network, we used a group of n healthy samples to calculate the Pearson correlation coefficient (PCC) of each pair of genes as background value of the corresponding edge, denoted as PCC_n. PCC_n is defined as follows:

$$PCC_n(x_1, x_2) = \frac{E(x_1x_2) - E(x_1)E(x_2)}{\sqrt{E(x_1^2) - E^2(x_1)}\sqrt{E(x_2^2) - E^2(x_2)}} \quad (1)$$

where x_1 and x_2 are the expression profiles of a pair of genes that correspond to an edge, and E represents the operator of mathematical expectation.

Next, a single disease sample was added to the set of the normal samples, and the new PCC was calculated and denoted as PCC_{n+1}. After that, the difference between background and interference values for the edges in each fully connected network could be quantified, which is represented as ΔPCC_n (equal to PCC_{n+1}–PCC_n). The difference was derived from the influence of the newly added disease sample, thus it can reflect the specific characteristics of this single sample. Statistically, ΔPCC_n obeys the volcano distribution. Therefore, the significance of ΔPCC_n can be estimated by the hypothesis test Z-test. Z-value is calculated as follows:

$$Z = \frac{\Delta PCC_n}{(1 - PCC_n^2)/(n-1)} \quad (2)$$

Statistical Difference of Gene Expressions Between Single Disease Sample and Normal Samples

The statistical difference of genes between single disease sample and normal samples in the expression level was calculated by fold change:

$$FC(x_i) = \frac{b}{\bar{a}} \quad (3)$$

where b represents the expression value of gene x_i in the individual disease sample and $\bar{a}$ is the mean of expression values of gene x_i over the n healthy samples.

Pathway Activation for a Single Sample

Based on the single sample analysis, we used AUCpath to estimate the activation of a pathway, which can evaluate the enrichment of an attention set as an area under the receiving operating characteristic curve (AUC) according to the ranking of all objects in a fully connected network. There were two sets, called the attention set and the background set. The attention set contained the subset of objects we considered as important, while the background set contained all the possible objects except important objects. We described the states of pathways from the aspect of genes and regulatory links.

From the perspective of edges, the input was the Z-value of all edges in each fully connected network, and the output was the activation of each pathway. The scoring approach was divided into two steps. First, the edges that exist in the pathway were regarded as an attention set (i.e., positive label), and the artificially added edges (in the step of the construction of fully connected network) were considered as the background set (i.e., negative label). Then, all edges in each fully connected network were ranked in ascending order of their Z-values. Second, AUC

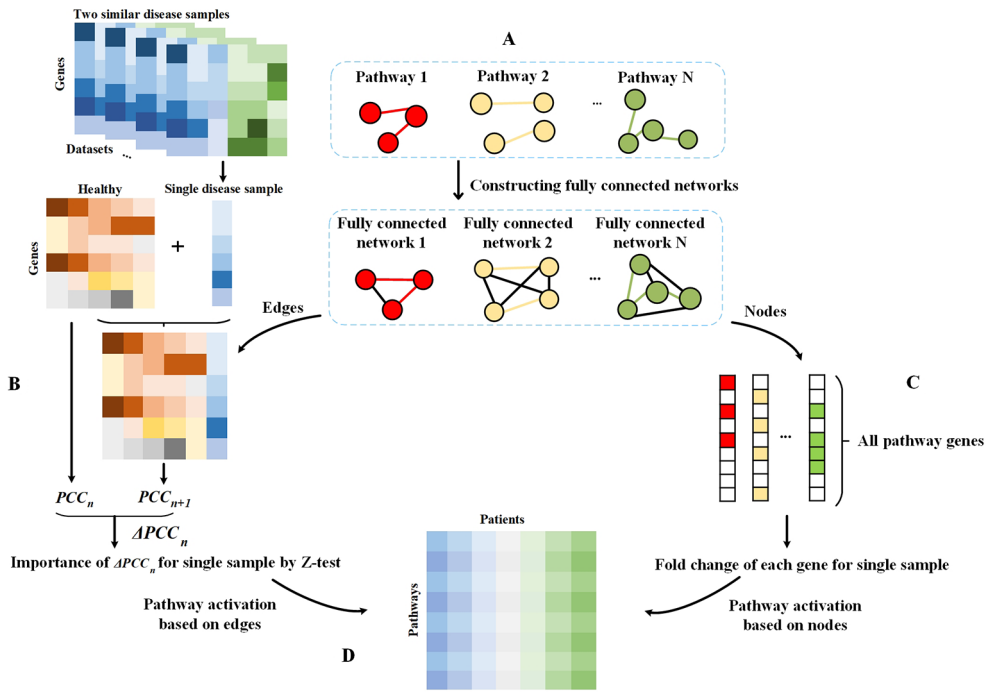

FIGURE 1 | Schematic diagram of the framework. **(A)** Fully connected networks are derived from pathways. The colored edges represent the real interactions in pathways and are regarded as attention sets, and the black edges are artificially added to construct fully connected networks and regarded as the background sets. **(B)** Single sample theory for the evaluation of differential value of each edge in fully connected networks. **(C)** Obtaining all genes in pathways. For a pathway, genes on this pathway are considered as the attention set and others are treated as the background set. The fold change value of each gene in each disease sample relative to the normal sample is evaluated for subsequent analysis. **(D)** PASS expression matrix. For each pathway, the AUCpath is used to evaluate the enrichment of edges in a pathway as an AUC according to the ranking of all edges in a fully connected network, whereby all edges were ranked according to their Z-scores. For each pathway, the AUCpath is used to assess the enrichment of genes in the pathway as an AUC based on the ranking of all pathway genes, whereby all genes were ranked according to the gene expression data.

was applied to evaluate whether edges in a pathway are enriched in the top ranking, and we thus regarded the AUC value as the quantitative indicator of pathway activation. It is defined as follows:

$$AUCpath = \frac{\sum_{i \in importantSubset} rank_i - \frac{m(1+m)}{2}}{m \times n} \quad (4)$$

where $rank_i$ represents the ranked position of the i-th edge of the attention set, m represents the number of edges in the attention set, and n is the number of edges in the background set.

Besides, considering that genes were also critical for mining effective information, we calculated the pathway activation from the perspective of genes. We first obtained all genes in pathways. For each pathway, genes on it were regarded as an attention set, and other genes were considered as the background set. Then, we assessed the enrichment of genes in the attention set as AUC based on the ranking of all genes, whereby all genes were ranked in ascending order according to their fold change between a single disease sample and normal samples.

After the evaluation of pathway activation from both nodes and edges, we obtained a matrix with PASS scores for pathways and patients.

RESULTS AND DISCUSSION

Stronger Effectiveness of PASS Compared to the Representative Feature Engineering Methods

We built a comprehensive scheme to demonstrate the performance of our approach for distinguishing two similar diseases as well as compare them with other state-of-the-art feature engineering methods. We selected seven representative methods from three aspects: network-based, GO-based and pathway-based methods, that is, NetRank (Winter et al., 2012), stSVM (Cun and Fröhlich, 2013), comparative network stratification (CNS) (Zhang et al., 2017), principal component analysis (PCA) (Young and Craft, 2016), normal tissue centroid (NTC) (Young and Craft, 2016), gene expression deviation (GED) (Young and Craft, 2016), and probabilistic pathway score (PROPS) (Han et al., 2017). For a better comparison, we downloaded the PPI network from STRING database (http://string-db.org/) for NetRank, stSVM and CNS, and collected biological processes (BP) terms of Gene Ontology (GO) (http://www.geneontology.org/) for CNS.

NetRank (Winter et al., 2012) is a modification of PageRank. For a given gene, NetRank identifies the rank of a gene according to the rank of its neighbors in a PPI network.

stSVM (Cun and Fröhlich, 2013) is a feature selection method which smooths the marginal statistic for differential expression genes by random walk kernel.

CNS (Zhang et al., 2017) is a framework that captures functional features for discriminating the disease states. Genes that are enriched by the same function (GO term) are aggregated through a flux balance model, and functional modules that maximize the distinction between UC and CD are then obtained.

For genes on each pathway, PCA (Young and Craft, 2016) compresses gene expression data and extracts principal components for the classification of disease status. For the hyperspace formed by genes on a particular pathway, NTC (Young and Craft, 2016) treats each disease sample as a point in the hyperspace and computes the Euclidean distance between the coordinates of disease samples and healthy samples. GED (Young and Craft, 2016) firstly uses the Kolmogorov–Smirnov test to capture genes that have the different distribution in normal and disease samples, and scores of those genes are then calculated based on the expression deviation in normal and disease samples. According to the scores, GED gives two features to each pathway, one for over-expression and one for under-expression. PROPS (Han et al., 2017) regards each pathway as a Gaussian Bayesian model. For each gene, after calculating the parameters in the model through normal samples, probabilistic pathway scores can be obtained using the loglikelihood values.

Improved Discrimination of PASS Evaluated by Classification Performance Analysis

We used the random forest classifier to verify the classification results and applied three-fold cross-validation considering the small sample size of several datasets. For unbiased evaluation, we repeated these experiments for a total of 500 times for the entire datasets. The results of eight methods are shown as ROC curves and AUC corresponding to the ROC in **Figure 2** and **Table 2**, respectively. Although the AUC of PROPS on GSE3365 somewhat exceeded PASS, and the AUC of PCA on GSE16879 was equal to PASS, our method was more stable and more prominent than the other seven methods on the five datasets.

Analysis of Differential Pathways With Significance According to PASS

In order to validate the effectiveness of PASS features, we analyzed the differential pathways according to the PASS index. The p-value was calculated using two-sample t-test for the five datasets. **Supplementary Figure S1** shows the quantitative distribution of p-value of differential pathways based on the PASS scores for the five datasets. The pathway activation we defined can acquire lots of differential features with significance in two similar diseases, which indicates that the PASS index can widen the gap between UC and CD.

We analyzed pathways that were differentially expressed (p-value < 0.05) on all the datasets (**Supplementary Table S1**). The majority of differential pathways have been shown to be related to IBD as reported in the literature (**Table 3**). These pathways not only demonstrate the metabolic and immune abnormalities of IBD, but they also reveal the pathogenesis of IBD from specific perspectives. Furthermore, the expression of genes in differential pathways related to IBD can reflect the changes in the course of disease. For the differential pathways associated with IBD, we analyzed the up-regulation and down-regulation of differentially expressed genes with significance in UC and normal samples, CD and normal samples. **Figure 3** shows the Venn diagrams of *Epstein-Barr virus infection* pathway, and others are shown in Supplementary **Figures S2–S10**. Most genes have the same regulatory relationship in UC and CD, but a small number of genes have different expressions. This also verifies that these two

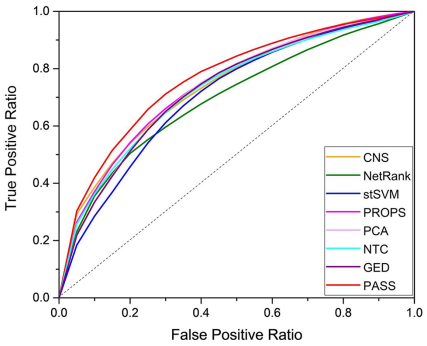

FIGURE 2 | Aggregate ROC curves.

TABLE 2 | Classification performance comparison on independent datasets.

Methods	GSE9686	GSE3365	GSE36807	GSE71730	GSE16879
PASS	0.94	0.77	0.78	0.74	0.72
NetRank	0.88	0.75	0.65	0.69	0.56
stSVM	0.88	0.72	0.75	0.71	0.55
CNS	0.91	0.75	0.75	0.70	0.69
PCA	0.91	0.66	0.73	0.69	0.72
NTC	0.89	0.72	0.75	0.67	0.68
GED	0.88	0.70	0.73	0.67	0.70
PROPS	0.88	0.78	0.70	0.73	0.67

TABLE 3 | Differential pathways related to IBD.

Entry	Name	Reference
hsa05169	Epstein-Barr virus infection	(Yanai et al., 1999)
hsa00190	Oxidative phosphorylation	(Soderholm et al., 2000; Söderholm et al., 2002)
hsa00531	Glycosaminoglycan degradation	(Lee et al., 2008b)
hsa00730	Thiamine metabolism	(Mehanna et al., 2008)
hsa00860	Porphyrin and chlorophyll metabolism	(Jansson et al., 2009)
hsa04012	ErbB signaling pathway	(Ando et al., 2013)
hsa04340	Hedgehog signaling pathway	(Ghorpade et al., 2013)
hsa04920	Adipocytokine signaling pathway	(Karmiris et al., 2006)
hsa00062	Fatty acid elongation	(Belluzzi et al., 2000)
hsa00020	Citrate cycle (TCA cycle)	(Schicho et al., 2012)

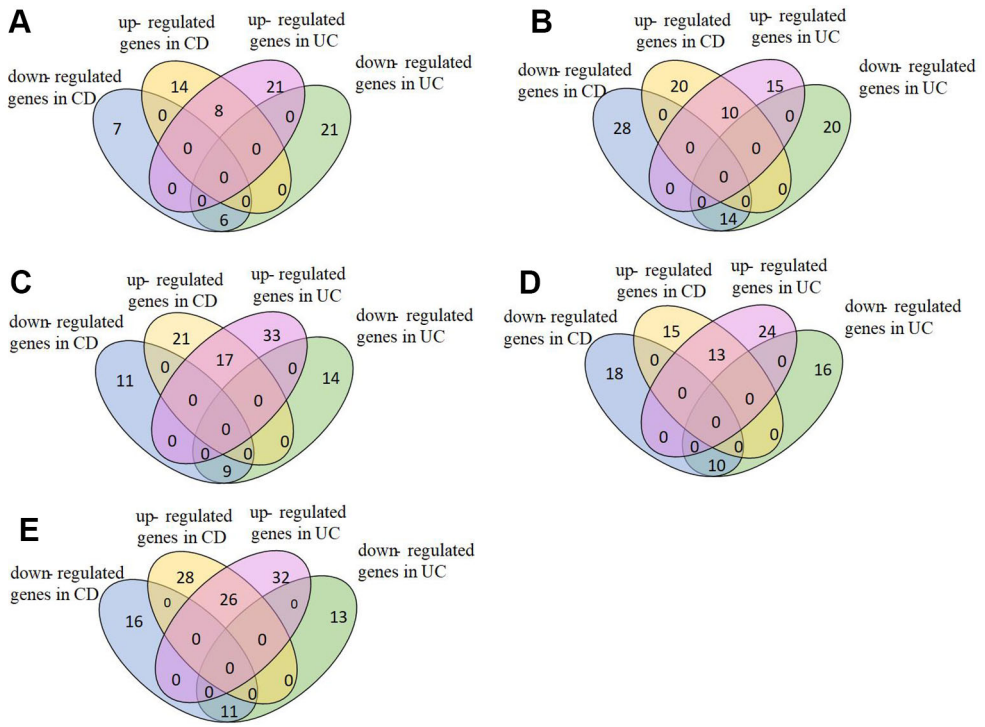

FIGURE 3 | Expression of genes in Epstein-Barr virus infection pathway. **(A)**GSE9686,**(B)**GSE3365, **(C)**GSE36807, **(D)**GSE71730, **(E)**GSE16879.

types of diseases are very similar, but there are differences between them.

Furthermore, we have visualized samples using the two principal components of our PASS features and overlaid the classification results from PASS model (**Figure 4**). The CD samples misclassified as UC and the UC samples misclassified as CD are mainly concentrated in the overlapping regions of the two types of diseases. However, some UC samples are more like CD samples, while some CD samples resemble UC samples, which leads to the misclassification of samples.

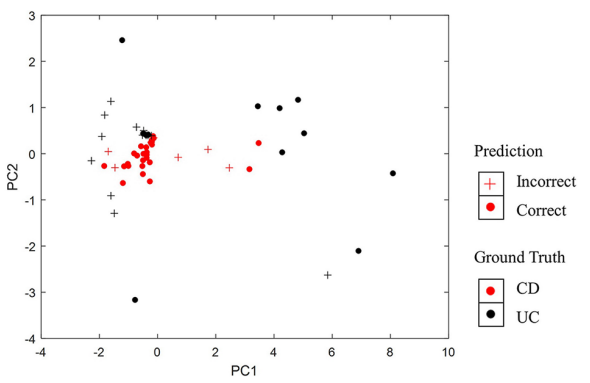

FIGURE 4 | Visualization of classification results using the two principal components of PASS features.

Enrichment of Known Disease-Associated Genes

After choosing a p-value < 0.01 as the threshold of statistical significance, we obtained the significant differential pathways. Next, we analyzed the enrichment of the known disease-associated genes (DAGs) in differential expression pathways. DAGs relevant to UC and CD were collected from DisGeNET (Piñero et al., 2016), and a hypergeometric test was used to calculate the p-value of the enrichment of DAGs:

$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \quad (5)$$

where N is the number of genes in all pathways, M is the number of DAGs, n is the number of genes in the differential pathways, and m is the number of DAGs enriched in the differential pathways.

For convenience, we transformed p-value to $-log_{10}(p\text{–}value)$. We compared the statistical significance of the enrichment of DAGs in the significant differential pathways identified by PASS index with other pathway-based indexes (**Figure 5**). It shows that, with the exception of being outperformed by PROPS in GSE36807, the differential pathways obtained from PASS values have the statistical significance of the enrichment of DAGs and

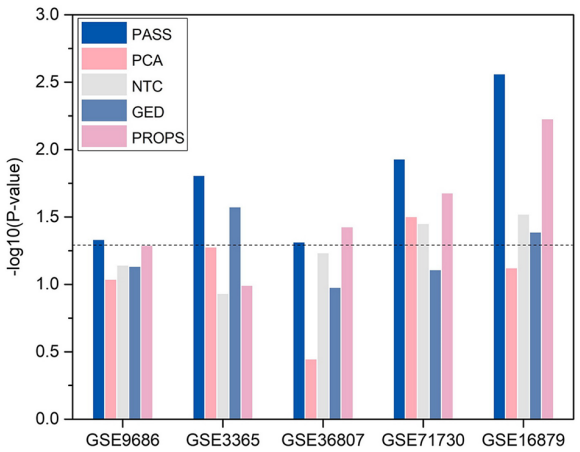

FIGURE 5 | Enrichment of known disease-associated genes in the significant differential pathways.

have lower p-values than other methods in all datasets. This indicates that the PASS index has the ability to identify differential features enriched by DAGs.

CONCLUSION

Complex diseases are not determined by a single gene, but by the combination of multiple genes, multiple factors, genetics, and the environment, similar complex diseases are more difficult to diagnose due to similar symptoms. In this study, we have presented PASS as a novel framework for classifying two main types of IBD from a single disease sample rather than a population of patients. For each pathway, we evaluated the difference between each patient and healthy sample from the perspective of genes and their interactions and calculated the pathway activation of individual samples. From the edge aspect, we constructed a fully connected network for each pathway, where edges in the pathway were regarded as the attention sets and artificially added edges were used as the background sets. Subsequently, we calculated the extent of perturbation of each edge based on single sample theory. From the node perspective, we collected all genes on all pathways. For each pathway, nodes on it were the attention set and others were the background set. Then, we evaluated the statistic difference of each node between single patient and healthy samples. Hereafter, we evaluated the pathway activation of each patient by computing the enrichment of attention set as an AUC according to the ranking of all genes or edges in the fully connected network.

We applied our method to UC and CD, which are two similar complex diseases of IBD. We compared PASS with seven state-of-the-art approaches (NetRank, stSVM, CNS, PCA, NTC, GED, and PROPS) on five IBD datasets. The results show that our PASS had the more discriminative power and was more stable than other seven methods. Besides, the PASS index can capture more differential expressed pathways with biological interpretability, which indicates that our PASS feature can widen the gap between UC and CD and aid researchers in comprehending the pathogenesis of these two similar complex diseases.

Our method can be applied to the classification of two similar diseases and has improved classification accuracy compared to seven state-of-the-art methods. However, due to the complexity and difficulty of similar complex diseases, there is still a space for improvement in the discriminative power. The performance of the PASS method relies on the all human pathway data and the topology of pathways, and more complete pathway information can better reveal the biological processes within cells and the statistic difference between a single disease sample and healthy samples calculated by our method can be also more accurate. With the rapid development of human interaction databases, we believe that the completer and more accurate pathway information could help to further improve the diagnosis of UC and CD.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at Gene Expression Omnibus (GSE9686, GSE3365, GSE36807, GSE71730, GSE16879).

AUTHOR CONTRIBUTIONS

XL and RZ conceived and designed the experiments. XL and XC performed the experiments and analyzed the data. XL wrote the paper. ML, JX, F-XW, and JW supervised the experiments and reviewed the manuscript.

FUNDING

This work was supported in part by the National Natural Science Foundation of China (61832019, 61702054), the 111 Project (No. B18059), the Hunan Provincial Innovation Foundation For Postgraduate (CX20190123), and the Hunan Provincial Natural Science Foundation of China (Grant No. 2018JJ3568).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01401/full#supplementary-material

REFERENCES

Akobeng, A. K., Zhang, D., Gordon, M., and MacDonald, J. K. (2016). Oral 5-aminosalicylic acid for maintenance of medically-induced remission in Crohn's disease. *Cochrane Database Syst. Rev.* 9, CD003715. doi: 10.1002/14651858.CD003715.pub3

Amar, D., Hait, T., Izraeli, S., and Shamir, R. (2015). Integrated analysis of numerous heterogeneous gene expression profiles for detecting robust disease-specific biomarkers and proposing drug targets. *Nucleic Acids Res.* 43, 7779–7789. doi: 10.1093/nar/gkv810

Ananthakrishnan, A. N. (2015). Epidemiology and risk factors for IBD. *Nat. Rev. Gastroenterol. Hepatol.* 12, 205. doi: 10.1038/nrgastro.2015.34

Ando, Y., Yang, G.-X., Kenny, T. P., Kawata, K., Zhang, W., Huang, W., et al. (2013). Overexpression of microRNA-21 is associated with elevated pro-inflammatory cytokines in dominant-negative TGF-β receptor type II mouse. *J. Autoimmun.* 41, 111–119. doi: 10.1016/j.jaut.2012.12.013

Arijs, I., De Hertogh, G., Lemaire, K., Quintens, R., Van Lommel, L., Van Steen, K., et al. (2009). Mucosal gene expression of antimicrobial peptides in inflammatory bowel disease before and after first infliximab treatment. *PloS One* 4, e7984. doi: 10.1371/journal.pone.0007984

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25. doi: 10.1038/75556

Baumgart, D. C., and Sandborn, W. J. (2007). Inflammatory bowel disease: clinical aspects and established and evolving therapies. *Lancet* 369, 1641–1657. doi: 10.1016/S0140-6736(07)60751-X

Belluzzi, A., Boschi, S., Brignola, C., Munarini, A., Cariani, G., and Miglio, F. (2000). Polyunsaturated fatty acids and inflammatory bowel disease. *Am. J. Clin. Nutr.* 71, 339s–342s. doi: 10.1093/ajcn/71.1.339s

Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., et al. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439, 353. doi: 10.1038/nature04296

Burczynski, M. E., Peterson, R. L., Twine, N. C., Zuberek, K. A., Brodeur, B. J., Casciotti, L., et al. (2006). Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. *J. Mol. Diagn.* 8, 51–61. doi: 10.2353/jmoldx.2006.050079

Carey, R., Jurickova, I., Ballard, E., Bonkowski, E., Han, X., Xu, H., et al. (2007). Activation of an IL-6: STAT3-dependent transcriptome in pediatric-onset inflammatory bowel disease. *Inflamm. Bowel Dis.* 14, 446–457. doi: 10.1002/ibd.20342

Choi, J., Park, S., Yoon, Y., and Ahn, J. (2017). Improved prediction of breast cancer outcome by identifying heterogeneous biomarkers. *Bioinformatics* 33, 3619–3626. doi: 10.1093/bioinformatics/btx487

Cun, Y., and Fröhlich, H. (2013). Network and data integration for biomarker signature discovery via network smoothed t-statistics. *PloS One* 8, e73074. doi: 10.1371/journal.pone.0073074

Drier, Y., Sheffer, M., and Domany, E. (2013). Pathway-based personalized analysis of cancer. *Proc. Natl. Acad. Sci.* 110, 6388–6393. doi: 10.1073/pnas.1219651110

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210. doi: 10.1093/nar/30.1.207

Ghorpade, D. S., Sinha, A. Y., Holla, S., Singh, V., and Balaji, K. N. (2013). NOD2-nitric oxide-responsive microRNA-146a activates Sonic hedgehog signaling to orchestrate inflammatory responses in murine model of inflammatory bowel disease. *J. Biol. Chem.* 288, 33037–33048. doi: 10.1074/jbc.M113.492496

Gurram, B., Salzman, N., Kaldunski, M., Jia, S., Li, B., Stephens, M., et al. (2016). Plasma-induced signatures reveal an extracellular milieu possessing an immunoregulatory bias in treatment-naive paediatric inflammatory bowel disease. *Clin. Exp. Immunol.* 184, 36–49. doi: 10.1111/cei.12753

Haider, S., Yao, C. Q., Sabine, V. S., Grzadkowski, M., Stimper, V., Starmans, M. H., et al. (2018). Pathway-based subnetworks enable cross-disease biomarker discovery. *Nat. Commun.* 9, 4746. doi: 10.1038/s41467-018-07021-3

Han, L., Maciejewski, M., Brockel, C., Gordon, W., Snapper, S. B., Korzenik, J. R., et al. (2017). A probabilistic pathway score (PROPS) for classification with applications to inflammatory bowel disease. *Bioinformatics* 34, 985–993. doi: 10.1093/bioinformatics/btx651

Hanauer, S. B. (2006). Inflammatory bowel disease: epidemiology, pathogenesis, and therapeutic opportunities. *Inflamm. Bowel Dis.* 12, S3–S9. doi: 10.1097/01.MIB.0000195385.19268.68

Huang, E., Ishida, S., Pittman, J., Dressman, H., Bild, A., Kloos, M., et al. (2003). Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nat. Genet.* 34, 226. doi: 10.1038/ng1167

Jansson, J., Willing, B., Lucio, M., Fekete, A., Dicksved, J., Halfvarson, J., et al. (2009). Metabolomics reveals metabolic biomarkers of Crohn's disease. *PloS One* 4, e6386. doi: 10.1371/journal.pone.0006386

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Karmiris, K., Koutroubakis, I. E., Xidakis, C., Polychronaki, M., Voudouri, T., and Kouroumalis, E. A. (2006). Circulating levels of leptin, adiponectin, resistin, and ghrelin in inflammatory bowel disease. *Inflamm. Bowel Dis.* 12, 100–105. doi: 10.1097/01.MIB.0000200345.38837.46

Lawrance, I. C., Fiocchi, C., and Chakravarti, S. (2001). Ulcerative colitis and Crohn's disease: distinctive gene expression profiles and novel susceptibility candidate genes. *Hum. Mol. Genet.* 10, 445–456. doi: 10.1093/hmg/10.5.445

Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T., and Lee, D. (2008a). Inferring pathway activity toward precise disease classification. *PloS Comput. Biol.* 4, e1000217. doi: 10.1371/journal.pcbi.1000217

Lee, H.-S., Han, S.-Y., Bae, E.-A., Huh, C.-S., Ahn, Y.-T., Lee, J.-H., et al. (2008b). Lactic acid bacteria inhibit proinflammatory cytokine expression and bacterial glycosaminoglycan degradation activity in dextran sulfate sodium-induced colitic mice. *Int. Immunopharmacol.* 8, 574–580. doi: 10.1016/j.intimp.2008.01.009

Liu, W., Li, C., Xu, Y., Yang, H., Yao, Q., Han, J., et al. (2013). Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. *Bioinformatics* 29, 2169–2177. doi: 10.1093/bioinformatics/btt373

Liu, X., Wang, Y., Ji, H., Aihara, K., and Chen, L. (2016). Personalized characterization of diseases using sample-specific networks. *Nucleic Acids Res.* 44, e164–e164. doi: 10.1093/nar/gkw772

Markert, E. K., Mizuno, H., Vazquez, A., and Levine, A. J. (2011). Molecular classification of prostate cancer using curated expression signatures. *Proc. Natl. Acad. Sci.* 108, 21276–21281. doi: 10.1073/pnas.1117029108

Mehanna, H. M., Moledina, J., and Travis, J. (2008). Refeeding syndrome: what it is, and how to prevent and treat it. *Bmj* 336, 1495–1498. doi: 10.1136/bmj.a301

Montero-Meléndez, T., Llor, X., García-Planella, E., Perretti, M., and Suárez, A. (2013). Identification of novel predictor classifiers for inflammatory bowel disease by gene expression profiling. *PloS One* 8, e76235. doi: 10.1371/journal.pone.0076235

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

Söderholm, J. D., Olaison, G., Peterson, K., Franzen, L., Lindmark, T., Wirén, M., et al. (2002). Augmented increase in tight junction permeability by luminal stimuli in the non-inflamed ileum of Crohn's disease. *Gut* 50, 307–313. doi: 10.1136/gut.50.3.307

Schicho, R., Shaykhutdinov, R., Ngo, J., Nazyrova, A., Schneider, C., Panaccione, R., et al. (2012). Quantitative metabolomic profiling of serum, plasma, and urine by 1H NMR spectroscopy discriminates between patients with inflammatory bowel disease and healthy individuals. *J. Proteome Res.* 11, 3344–3357. doi: 10.1021/pr300139q

Soderholm, J. D., Wiren, M., Franzen, L. E., Perdue, M. H., and Olaison, G. (2000). Topical phase effects of acetylsalicylic acid on human small bowel epithelium: Inhibition of oxidative phosphorylation and increased tight junction permeability. *Gastroenterology* 118, A811. doi: 10.1016/S0016-5085(00)85386-X

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Thomas, D. C., Baurley, J. W., Brown, E. E., Figueiredo, J. C., Goldstein, A., Hazra, A., et al. (2008). Approaches to complex pathways in molecular epidemiology: summary of a special conference of the American Association for Cancer Research. *Cancer Res.* 68, 10028–10030. doi: 10.1158/0008-5472.CAN-08-1690

Winter, C., Kristiansen, G., Kersting, S., Roy, J., Aust, D., Knösel, T., et al. (2012). Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PloS Comput. Biol.* 8, e1002511. doi: 10.1371/journal.pcbi.1002511

Wu, F., Dassopoulos, T., Cope, L., Maitra, A., Brant, S. R., Harris, M. L., et al. (2007). Genome-wide gene expression differences in Crohn's disease and

ulcerative colitis from endoscopic pinch biopsies: insights into distinctive pathogenesis. *Inflamm. Bowel Dis.* 13, 807–821. doi: 10.1002/ibd.20110

Yanai, H., Shimizu, N., Nagasaki, S., Mitani, N., and Okita, K. (1999). Epstein-Barr virus infection of the colon with inflammatory bowel disease. *Am. J. Gastroenterol.* 94, 1582. doi: 10.1111/j.1572-0241.1999.01148.x

Young, M. R., and Craft, D. L. (2016). Pathway-informed classification system (PICS) for cancer analysis using gene expression data. *Cancer Inf.* 15, 151–161. CIN.S40088. doi: 10.4137/CIN.S40088

Yousefi, M. R., and Dougherty, E. R. (2012). Performance reproducibility index for classification. *Bioinformatics* 28, 2824–2833. doi: 10.1093/bioinformatics/bts509

Zhang, J. D., and Wiemann, S. (2009). KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics* 25, 1470–1471. doi: 10.1093/bioinformatics/btp167

Zhang, C., Liu, J., Shi, Q., Zeng, T., and Chen, L. (2017). Comparative network stratification analysis for identifying functional interpretable network biomarkers. *BMC Bioinf.* 18, 48. doi: 10.1186/s12859-017-1462-x

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.