Identification of key genes and pathways associated with Crohn's disease by bioinformatics analysis

Zheng Wang¹, Jie Zhu¹, Lixian Ma¹

¹Department of Infectious Diseases, Qilu Hospital, Shandong University,
Shandong, China;

Corresponding author: Lixian Ma, E-mail: mlx_sdu@163.com

Abstract

Crohn's disease is a type of inflammatory bowel disease posing a significant threat to human health all over the world. Genome-wide gene expression profiles of mucosal colonic biopsies have provided some insight into the molecular mechanisms of Crohn's disease. However, the exact pathogenesis is unclear. This study aimed to identify key genes and significant signaling pathways associated with Crohn's disease by bioinformatics analysis. To identify key genes, an integrated analysis of gene expression signature was conducted using a robust rank aggregation approach. A total of 179 Crohn's disease patients and 94 healthy controls from nine public microarray datasets were included. MMP1 and CLDN8 were two key genes screened from the differentially expressed genes. Connectivity Map predicted several small molecules as possible adjuvant drugs to treat CD. Besides, we used weighted gene co-expression network analysis to explore the co-expression modules associated with Crohn's disease pathogenesis. Seven main functional modules were identified, of which black module showed the highest correlation with Crohn's disease. The genes in black module mainly enriched in Interferon Signaling and defense response to virus. Blue module was another important module and enriched in several signaling pathways, including extracellular matrix organization, inflammatory response and blood vessel development. There were also several other

meaningful functional modules which enriched in many biological processes. The present study identified a number of key genes and pathways correlated with Crohn's disease and potential drugs to combat it, which might offer insights into Crohn's disease pathogenesis and provide a clue to potential treatments.

Introduction

Crohn's disease (CD) is a chronic nonspecific inflammatory bowel disease (IBD), which may affect any region of the gastrointestinal tract intermittently with the terminal ileum and colon being the most common [1]. The incidence and prevalence of CD were highest in developed countries, with the highest annual incidence in North America (20.2 per 100,000) and Europe (12.7 per 100,000). Similarly, the highest worldwide prevalence of CD was found in Europe (322 per 100,000) and North America (319 per 100,000) [2]. Furthermore, the incidence of CD in the developing countries continues to increase. Regarding sexual distribution, there was no sex preponderance in adult CD [1]. About 12% of CD patients have a family history of IBD, with an even higher proportion of family cases in younger individuals [3]. Although the exact etiology of CD remains uncertain to date, there is evidence that it involves a complex interaction between the individual's genetic susceptibility, external environment, intestinal microbial flora and the immune responses [4, 5]. Previous genome-wide association studies have reported numerous loci associated with CD, which included NOD2, ATG16L1 and IRGM. NOD2 was identified as the first susceptibility gene for CD [6]. NOD2, a cytosolic pattern recognition receptor for the bacterial proteoglycan fragment muramyl dipeptide, plays a significant role in maintaining immunological homeostasis in the intestine. NOD2

mutations are genetically correlated with an increased risk for the development of CD [5, 7]. Moreover, genetic analyses have identified autophagy genes ATG16L1 and IRGM genes as susceptibility genes for CD [8]. Together, these common susceptibility genes account for 13.1% of variance in disease liability in CD [9].

A large amount of gene expression datasets are currently freely available from the Gene Expression Omnibus (GEO) database. Due to the small sample sizes in a single dataset and discrepancies of the characteristics within multiple heterogeneous datasets, it is necessary to integrate those massive datasets through systems biology tools and finally receive the stable and credible results [10]. So far, comprehensive integrated analysis of gene expression datasets in CD is still missing. Therefore, we executed integrated analysis method by using robust rank aggregation (RRA) method to detect the consistently significant differentially expressed genes (DEGs) from various expression profiling datasets [11]. In addition, we performed weighted gene co-expression network analysis (WGCNA) to categorise those DEGs into a number of functional modules, and followed by functional and pathway enrichment analysis [12]. Connectivity Map (CMap) represents a useful bioinformatic technique for establishing the connections among genes, drugs and diseases [13, 14]. CMap has been applied to discover drug action mechanisms and disease pathogenesis, as well as to assist in the

identification of novel potential therapeutics [15, 16]. The DEGs of CD and control intestinal tissues were used to query the CMap database, which includes more than 7000 expression profiles corresponding to 1309 drugs. In the current study, we used CMap to explore therapeutic drugs that could potentially be effective against CD.

Here, we screened the differentially expressed genes from the intestinal tissue microarray data of CD patients and healthy people by means of bioinformatics technology, and obtained the signal pathways related to the disease through enrichment analysis. Moreover, the potential therapeutic drugs for CD were explored via the CMap database. The study will help to understand the pathogenesis of CD and lay a theoretical foundation for molecular diagnosis and targeted therapy.

Methods

Data Collection and eligibility criteria

The Probe signal data from the key word “ Crohn's disease biopsy” were downloaded from Gene Expression Omnibus (GEO) database(<http://www.ncbi.nlm.nih.gov/geo/>). The selection criteria for this study were displayed as follows: (a) gene expression datasets which included gene microarray chip technology; (b) studies comparing gene expressions between active crohn's disease and normal controls tissue in human samples; (c) the number of samples in each gene expression profiling dataset was no less than 6; (d) Processed data or raw data can be used for reanalysis should be provided in these databases. Studies not meeting these selection criteria were not remained in the analysis.

RRA analysis

Data preprocessing steps and statistical analysis were performed using R software(<https://www.r-project.org/>) and Bioconductor packages [17]. Microarray datasets were normalized using RMA [18]and differential expression analysis was implemented using the limma R package [19]. The statistical operations were performed using the Robust Rank Aggreg package of R statistical software [11]. Genes were ranked by statistical significance which was evaluated by the expression fold change level and P values.

Screening of small drug molecules

To explore potential therapeutic drugs against CD, the DEGs were compared to the expression signatures of 1309 drugs in the CMap database (<https://portals.broadinstitute.org/cmap/>) [13, 14]. The upregulated and downregulated DEGs generated through the RRA method that had a $|\log FC| > 0.5$ were imported into the CMap database for analysis. Compounds with negative connectivity scores, which indicated that the corresponding drug may reverse the expression of the query signature in CD, were recorded as potential therapeutic drugs. The small drug molecules with an average connectivity scores < -0.25 and P value < 0.05 were recorded (Table 2).

WGCNA

Gene co-expression network was implemented by the WGCNA package in R [12]. GSE16879 which including 37 CD patients and 12 controls was used to construct the WGCNA analysis. To include a sufficient amount of genes into WGCNA, genes with $p < 0.05$ and $|\log_{10} \text{fold changes}| (|\log FC|) > 0.14$ were chosen from the ranked gene list. When the soft threshold was defined as 11, the scale-free topology index was > 0.9 . In the present study, we chose a minimum module size of 150 for the genes dendrogram and the minimum height for merging modules at

0.15. The modules were randomly color-labeled.

Functional enrichment analysis

Meaningful co-expression modules were screened out for functional enrichment analysis, performing by metascape (<http://metascape.org>) [20]. Top ten clusters with their representative enriched terms were selected.

Statistical analysis

Statistical analyses were performed using R software version 3.5.0. $P < 0.05$ was considered to be statistically significant.

Results

CD microarray datasets

According to the screening criteria, 9 datasets in CD were finally included in this study: GSE10714 [21], GSE16879 [22], GSE36807 [23], GSE9686 [24], GSE10616 [25], GSE75214 [26], GSE6731 [27], GSE52746 [28], GSE68570 [29] (Table1). The description of the studies was provided in Table 1, such as GSE number, authors and number of samples. The number of CD patients in these studies ranged from 4 to 59, and the number of controls ranged from 3 to 22. A total of 179 CD patients and 94 healthy controls were finally included in this study.

Table 1 Summary of those 9 genome-wide gene expression datasets involving CD patients

	GSE number	Samples	Source types	Platform	authors
1	GSE10714	4 CD patients and 3 controls	colonic biopsies	GPL570	Galamb O, Sipos F, Solymosi N, Spisák S et al.
2	GSE16879	37 CD patients and 12 controls	colonic biopsies and ileal mucosal biopsy	GPL570	Arijs I, De Hertogh G, Lemaire K, Quintens R et al.
3	GSE36807	13 CD patients and 7 controls	colon biopsies	GPL570	Montero-Meléndez T, Llor X, García-Planella E, Perretti M et al.
4	GSE9686	11 CD patients and 8 controls	colon biopsies	GPL5760	Carey R, Jurickova I, Ballard E, Bonkowski E et al.
5	GSE10616	32 CD	colon biopsies	GPL576	Kugathasan S,

		patients and 16 controls		0	Baldassano RN, Bradfield JP, Sleiman PM et al.
6	GSE75214	59 CD patients and 22 controls	colonic biopsies and ileal mucosal biopsy	GPL624 4	Vancamelbeke M, Vanuytsel T, Farré R, Verstockt S et al.
7	GSE6731	7 CD patients and 4 controls	colonic biopsies	GPL830 0	Wu F, Dassopoulos T, Cope L, Maitra A et al.
8	GSE52746	10 CD patients and 17 controls	colonic biopsies	GPL179 96	Leal RF, Planell N, Kajekar R, Lozano JJ et al.
9	GSE68570	6 CD patients and 5 controls	colonic biopsies	GPL179 96	Leal RF, Planell N, Kajekar R, Lozano JJ et al.

205

206 RRA analysis

207 By using a RRA method to integrate 9 datasets, upregulated and
 208 downregulated ranked gene lists were obtained. S1 Table showed those
 209 top 150 upregulated and top 23 downregulated genes in CD patients.
 210 These top 150 upregulated genes had a $p < 7.4E^{-8}$ and a $|\log FC| > 1$.
 211 These top 23 downregulated genes had a $p < 2.1E^{-10}$ and a $|\log FC| > 1$.
 212 The top 20 upregulated and downregulated genes in the integrated
 213 microarray analysis were showed in Fig 1. Among these top genes, the
 214 abnormal expression of some genes had been well confirmed, such as
 215 MMP1, CLDN8 [30, 31]. However, the roles of other genes in CD have
 216 not been previously reported in the literature, such as PCK1, CHP2 .

217 **Fig 1. Top 20 upregulated and top 20 downregulated genes in CD.** Each

column represents one dataset and each row represents one gene.

Numbers in each rectangle show the logarithmic fold change of genes in

each dataset. Green indicates decreased gene expression and red

indicates increased gene expression.

Small drug molecule screening

The related small molecules with highly significant correlations were

listed in Table 2. Among these molecules, tetracycline showed minimum

p value and higher negative correlation, indicating the potential

molecular signature that may counteract those of CD.

Table 2 Results of CMap analysis

cmap name	mean	enrichment t	p
tetracycline	-0.41	-0.9	0.00004
methoxamine	-0.282	-0.835	0.00135
tubocurarine chloride	-0.542	-0.83	0.00151
Prestwick-675	-0.265	-0.818	0.00203
benzbromaro ne	-0.301	-0.874	0.00403
methocarbam ol	-0.28	-0.831	0.00959
colistin	-0.266	-0.724	0.0118
disopyramide	-0.348	-0.714	0.01361

amantadine	-0.349	-0.682	0.02222
pepstatin	-0.274	-0.638	0.04144

230

231 **WGCNA**

232 The expressions of 5666 DEGs generated from the RRA analysis were
 233 used for WGCNA, all of which had a $p < 0.05$ and a $|\log FC|$ of > 0.14 . As
 234 shown in Fig 2A, a power value of 11 was selected as the
 235 soft-thresholding to ensure a scale-free network. We set 150 as the least
 236 gene number of each gene network and 0.15 as the threshold for the
 237 merge of similar module. The WGCNA elucidated 7 coexpressed modules
 238 ranging in size from 269 to 1730 and the grey module represented a
 239 gene set that was not assigned to any of the modules (Fig 2B).
 240 Interactions of the 7 co-expression modules were analyzed (Fig 2C). The
 241 genes in each module were listed in S2 Table.

242

243 **Fig 2 Plots in the WGCNA analysis using gene expressions in 37 CD**
 244 **patients and 12 controls from GSE16879 datasets.** (A) Network topology
 245 of different soft-thresholding powers. The left panel displayed the
 246 influence of soft-threshold power on scale-free topology fit index; the
 247 right panel showed the influence of soft-threshold power on the mean
 248 connectivity. (B) Cluster dendrogram of coexpression genes and
 249 functional modules in CD. 7 co-expression modules were constructed

and was shown in different color. (C) Visualizing the gene network using a heatmap plot. Different colors of horizontal axis and vertical axis represented different modules. Light color represents low overlap and gradually dark red represents higher overlap.

Identifying the modules most correlated with clinical features has a great biological significance. As shown in Fig 3, black, blue, red and yellow modules positively correlated to CD ($p < 0.05$). Turquoise and brown modules negatively correlated to CD ($p < 0.05$). Brown and grey modules were positively correlated with a better clinical response to infliximab ($p < 0.05$) while blue and yellow modules was correlated negatively to treatment. Combined with Fig 4, we found that these six modules yielded two main clusters; one contained four modules (black, blue, red and yellow module) while the other contained two modules (brown and turquoise module). Moreover, two pairs of module combination had much higher adjacency degree and they are blue module and yellow module, black module and red module respectively.

The black module showed the highest correlation with CD ($r = 0.71$, $p = 1e-8$, Fig 3). The genes in black module mainly enriched in Interferon Signaling and defense response to virus. Genes in red module were enriched in Cell Cycle, cell division and cell cycle phase transition. Genes in blue module were enriched in several signaling pathways, including

extracellular matrix organization, inflammatory response, blood vessel development and leukocyte migration. The yellow module was similar to the blue module, mainly enriched lymphocyte activation, adaptive immune response, leukocyte activation involved in immune response. Some other signaling pathways were observed to enrich in other modules, such as monocarboxylic acid metabolic process, cofactor metabolic process and Metabolism of lipids in turquoise, Transport of small molecules in yellow. The main findings for the functional enrichment analysis were shown in Table 3.

Table 3 Pathway and process enrichment analysis of those functional coexpression modules in CD

Modules	GO	Category	Description	Count	%	Log10(P)	Log10(q)
Black module	R-HSA-91353	Reactome	Interferon	36	13.9534883721	-32.6634376812	-28.3520653194
	1	Gene Sets	Signaling				
	GO:0051607	GO Biological Processes	defense response to virus	29	11.2403100775	-21.6512199059	-18.4190287902
	GO:0002476	GO Biological Processes	antigen processing and presentation of endogenous peptide antigen via MHC class Ib	7	2.71317829457	-12.8199138524	-9.68463274966
	R-HSA-98316	Reactome	Class I MHC mediated antigen processing & presentation	24	9.3023255814	-11.3182005651	-8.38063594571
	9	Gene Sets					
	GO:0001817	GO Biological Processes	regulation of cytokine	29	11.2403100775	-9.93175256791	-7.09750146082

		production				
	GO:0045088	GO Biological Processes	regulation of innate immune response	19	7.36434108527	-7.5778005841 -4.91964073607
	GO:0032496	GO Biological Processes	response to lipopolysaccharide	17	6.58914728682	-7.10900753576 -4.49660517829
	M161	Canonical Pathways	PID IFNG PATHWAY	7	2.71317829457	-6.58416324878 -4.03473422999
	GO:0060330	GO Biological Processes	regulation of response to interferon-gamma	6	2.32558139535	-6.58267859823 -4.03473422999
	GO:0036498	GO Biological Processes	IRE1-mediated unfolded protein response	8	3.1007751938	-6.43708025362 -3.91809958131
	GO:0032787	GO Biological Processes	monocarboxylic acid metabolic process	123	7.33890214797	-28.1238108256 -23.8124384638
	GO:0051186	GO Biological Processes	cofactor metabolic process	115	6.861575179	-25.2673411673 -21.2569988012
	R-HSA-556833	Reactome Gene Sets	Metabolism of lipids	134	7.99522673031	-23.0687643444 -19.456361987
	GO:0044283	GO Biological Processes	small molecule biosynthetic process	125	7.45823389021	-19.2949845842 -16.0356832879
Turquoise module	GO:0017144	GO Biological Processes	drug metabolic process	130	7.75656324582	-19.0211189435 -15.8858378408
	GO:0006820	GO Biological Processes	anion transport	107	6.38424821002	-18.265073659 -15.2089738023
	GO:0008202	GO Biological Processes	steroid metabolic process	70	4.17661097852	-16.9725530488 -13.9833999817
	GO:0006091	GO Biological Processes	generation of precursor metabolites and energy	87	5.19093078759	-16.1087777292 -13.1398280482
	GO:0005975	GO Biological Processes	carbohydrate metabolic process	96	5.72792362768	-14.3452501064 -11.5109989993
	R-HSA-211859	Reactome Gene Sets	Biological oxidations	51	3.04295942721	-13.0028239955 -10.2825162407
	GO:0030198	GO Biological Processes	extracellular matrix organization	116	8.41189267585	-55.1296356537 -50.8182632918
Blue module	GO:0006954	GO Biological Processes	inflammatory response	149	10.8049311095	-38.8989948741 -35.1896825036

	GO:0001568	GO Biological Processes	blood vessel development	143	10.3698332125	-35.4750329051	-31.8626305477
	GO:0050900	GO Biological Processes	leukocyte migration	106	7.68672951414	-32.8831539596	-29.4748715848
	M5885	Canonical Pathways	NABA MATRISOME ASSOCIATED	131	9.49963741842	-28.6809763948	-25.448785279
	GO:0030155	GO Biological Processes	regulation of cell adhesion	116	8.41189267585	-26.555324351	-23.4200432482
	GO:0001816	GO Biological Processes	cytokine production	120	8.70195794054	-25.7212283729	-22.6139759938
	GO:0002274	GO Biological Processes	myeloid leukocyte activation	111	8.049311095	-24.9327449	-21.8518214596
	GO:0009611	GO Biological Processes	response to wounding	114	8.26686004351	-24.8377171008	-21.7816172441
	R-HSA-1474228	Reactome Gene Sets	Degradation of the extracellular matrix	49	3.55329949239	-24.5416313487	-21.5090125878
Red module	R-HSA-1640170	Reactome Gene Sets	Cell Cycle	97	24.1895261845	-62.4943601224	-58.1829877606
	GO:0051301	GO Biological Processes	cell division	73	18.2044887781	-40.4973015535	-36.6630504464
	GO:0044770	GO Biological Processes	cell cycle phase transition	61	15.2119700748	-30.4219724816	-27.0136901068
	GO:0006260	GO Biological Processes	DNA replication	36	8.97755610973	-21.1804243358	-18.23077981
	R-HSA-69306	Reactome Gene Sets	DNA Replication	26	6.48379052369	-19.9879102614	-17.123695931
	GO:0006281	GO Biological Processes	DNA repair	45	11.2219451372	-17.2330094157	-14.489838778
	R-HSA-69275	Reactome Gene Sets	G2/M Transition	28	6.9825436409	-17.1069967286	-14.3866889739
	GO:0051321	GO Biological Processes	meiotic cell cycle	29	7.2319201995	-15.6001764778	-12.9877741204
	GO:0071103	GO Biological Processes	DNA conformation change	30	7.4812967581	-15.1235789271	-12.5525692548
	hsa04110	KEGG Pathway	Cell cycle	21	5.23690773067	-14.4868416002	-11.94632125
Yellow module	GO:0046649	GO Biological Processes	lymphocyte activation	80	15.2671755725	-32.6589516941	-28.3475793323
	GO:0002250	GO Biological Processes	adaptive immune response	51	9.73282442748	-15.1514204069	-11.9539913974
	GO:0002366	GO Biological Processes	leukocyte activation involved in immune	52	9.92366412214	-13.4023570595	-10.4132039925

		response				
	GO:0001816	GO Biological Processes	cytokine production	52	9.92366412214	-13.3283280319 -10.3593783509
	GO:0002253	GO Biological Processes	activation of immune response	47	8.96946564885	-11.9744267874 -9.09441818977
	GO:0042113	GO Biological Processes	B cell activation	31	5.91603053435	-11.6617184904 -8.79750415991
	hsa04670	KEGG Pathway	Leukocyte transendothelial migration	19	3.62595419847	-10.9890453658 -8.22174104834
	GO:0046631	GO Biological Processes	alpha-beta T cell activation	19	3.62595419847	-10.1411382447 -7.44254973964
	R-HSA-1428517	Reactome Gene Sets	The citric acid (TCA) cycle and respiratory electron transport	21	4.00763358779	-9.30483909218 -6.64667924415
	GO:0050900	GO Biological Processes	leukocyte migration	35	6.67938931298	-9.07542858275 -6.47162639704
Brown module	R-HSA-382551	Reactome Gene Sets	Transport of small molecules	59	7.60309278351	-9.5739170205 -5.26254465869
	GO:0006091	GO Biological Processes	generation of precursor metabolites and energy	41	5.28350515464	-7.8508304058 -3.85022179525
	GO:0006820	GO Biological Processes	anion transport	47	6.05670103093	-7.36872884322 -3.75632648575
	GO:0006644	GO Biological Processes	phospholipid metabolic process	40	5.15463917526	-7.25271548512 -3.7194943737
	GO:0044283	GO Biological Processes	small molecule biosynthetic process	53	6.82989690722	-6.97639301874 -3.51011869695
	GO:0030029	GO Biological Processes	actin filament-based process	51	6.57216494845	-6.31734872826 -3.11311753775
	GO:0006520	GO Biological Processes	cellular amino acid metabolic process	30	3.86597938144	-6.24865452065 -3.08445390015
	GO:0048194	GO Biological Processes	Golgi vesicle budding	14	1.80412371134	-6.16109071399 -3.05383833484
	GO:0045216	GO Biological Processes	cell-cell junction organization	19	2.44845360825	-6.0978562781 -3.01693283768
	GO:0002399	GO Biological Processes	MHC class II protein complex assembly	46	0.51546391752	-5.90949586892 -2.87687710807

Discussion

Crohn's disease is a chronic inflammatory disorder which usually Lead to pathological changes at different locations along the length of gastrointestinal tract. The precise cause of CD has not been determined. In this study, we integrated the gene expression data to shed light on the pathological mechanism of CD.

A large number of remarkably upregulated and downregulated genes were identified through RRA analysis, some of which have been reported in previous literature in CD, such as MMP1, MMP3, REG1B ,REG1A, CLDN8, GUCA2A. Previous studies have reported that the expression of MMP1 and MMP3 were upregulated in the colonic mucosa with CD compared with control [32]. MMP1 and MMP3 belong to the family of metal dependent enzymes, which are capable of degrading a wide range of extracellular matrix components [33]. It is of note that MMP1 was involved in inflammation and have been implicated in tissue injury and repairing processes in CD [34]. In accordance with previous reports, the present study found that MMP-1 and MMP-3 were located in the blue module, which was mainly enriched in extracellular matrix organization and inflammatory response. Because MMP-1 and MMP-3 are the top two of upregulated differentially expressed genes ranking, over expression of MMP-1 and MMP-3 are probably key factors in the pathogenesis of CD. It was reported previously that REG1B and REG1A

were highly expressed in CD colonic mucosa [35]. However, the molecular mechanism underlying how REG1B and REG1A exert their role in CD has not been determined yet. CLDN8 is a member of the family of claudins, which mainly located in the cell membrane and associated with tight junctions of cell adhesion [36]. It has been reported that CLDN8 might play a significant role in the injury of intestinal epithelial barrier of CD [37]. The expression level of CLDN8 was remarkably downregulated in CD colonic mucosa compared to normal controls, but the molecular mechanism underlying its role remains unclear. As previously demonstrated, GUCA2A expression was markedly decreased in CD, but its molecular mechanism is poorly understood [38]. The other genes (such as PCK1,CHP2) are novel gene signatures of CD, and there is still lack of researches on its role in CD pathogenesis. Therefore, their aberrantly expressions need to be validated in future researches.

Several small molecules with potential therapeutic efficacy against CD were identified according to the CMap database. As a result of our screening effort, tetracycline has been identified as a potential therapeutic target for CD. Tetracycline were discovered in the 1940s and showed broad spectrum inhibitory activity against a wide range of microorganisms. Previous findings indicated that tetracycline exert pleiotropic functions independent of their antimicrobial activity, which include inhibition of MMPs, angiogenesis and inflammation [39]. Owing

to the ability to inhibit MMPs expression and treat overgrowth of bacteria in the small intestine, tetracycline may be prescribed to treat CD. We might suppose that tetracycline could play certain roles to combat CD; however, further investigation on a large population is required to firmly establish the therapeutic effect in CD.

The WGCNA which is a kind of bioinformatics analysis method was conducted to explore the molecular mechanisms of various disease [40, 41]. Several reports involved IBD had been studied by this method [42, 43]. In this study, 5666 CD associated genes generated from RRA analysis were used in WGCNA and most of them were classified into 7 co-expression modules. Six of the modules have significant associations with disease (Fig 3), which can be enriched more refined in three cluster (Fig 4), highlighting some new insights into the pathogenesis of CD. To further understand the importance of those modules in the pathogenesis of CD, enrichment analysis was performed using metaspice. The two most critical modules, black module and blue module, which have the most significant relationship with disease need to be addressed. Black module was mainly enriched in some biological processes, the most important of which were Interferon Signaling, defense response to virus (Table 3). It is generally known that IFN-gamma plays a key role in the early steps of installation of inflammation, promoting monocyte recruitment and activation, and

inducing the expression of other inflammatory cytokines. Interferon Signaling has been identified as a central aspect of innate immune response which induces a wide variety of antiviral proteins against pathogens infection [44, 45]. Furthermore, previous studies have shown that Epstein-Barr virus and cytomegalovirus were the two main viruses that are positively correlated with the onset of CD [46]. In addition, bacteriophage was other viral agent that has been suspected to play a role in CD pathogenesis [47]. Although there was not enough evidence that the virus is directly involved in CD, the study by Cadwell et al. [48] suggested that the combination of host genetic susceptibility and the presence of viral factors could lead to CD occurrence. What's more, several studies have found that for some CD patients with steroid resistance, antiviral therapy may be effective for relieving symptoms and improving the sensitivity of anti-inflammatory drugs [49]. From what have been discussed above, we may get the hypothesis that for some CD patients, especially those with steroid-refractory CD, virus infection and the sequential Interferon Signaling pathway may be the key factor in the initial stage of disease onset. Similarly, the enrichment analysis of genes in the salmon module mainly involved in extracellular matrix organization, inflammatory response and blood vessel development. Meanwhile, yellow module clustered with blue module mainly enriched in the pathway of lymphocyte activation. The pathways enriched from

the two important modules were all belong to the intermediate link of the inflammatory response process, which explains the general pathogenesis of CD, an autoimmune disease. Moreover, extracellular matrix organization and blood vessel development also play a key role in the process of intestinal fibrosis at the late stage of CD [50]. Intestinal fibrosis continues to develop after intestinal primary injury and inflammatory response are subsided [51]. As we know, severe intestinal fibrosis is the main cause of intestinal resection, which seriously affects the prognosis of patients. In other words, extracellular matrix organization, inflammatory response and blood vessel development are key factors in the development of CD.

Fig 3. Heatmap of the correlation between module eigengenes and clinical traits of CD. The table is color coded by correlation according to the The table was color coded by correlation according to the color legend Each cell contains the corresponding correlation and p-value

Fig 4. Heatmap plot of the adjacencies of modules. The change of color from blue (0) to red (1) in the heatmap represented the connectivity degree of different modules from weak to strong.

Conclusion

In conclusion, this study identified a number of key genes and co-expression biologically functional modules which may be involved in the progress of CD, based on differentially expressed genes between CD samples and normal mucosas. Additionally, the present study screened a small drug molecule named tetracycline, which might be exploited as adjuvant drug for improved therapeutics for CD. Our findings suggest that virus infectious may be the initial link in CD. Moreover, extracellular matrix organization, inflammatory response and blood vessel development are the key pathways for the development of CD, which not only participate in the stage of inflammatory response process, but also play a key role in the late phase of intestinal fibrosis. This point may provide a new idea for improving the prognosis of CD.

Acknowledgments

This work did not receive any specific funding from any sources.

Author Contributions

Conceptualization: Zheng Wang ,Jie Zhu, Lixian Ma

Data curation: Zheng Wang, Jie Zhu

Formal analysis: Zheng Wang, Jie Zhu, Lixian Ma

Investigation: Zheng Wang, Jie Zhu, Lixian Ma

Methodology: Zheng Wang

Software: Zheng Wang

Supervision: Lixian Ma

Visualization: Zheng Wang

Writing-original draft: Zheng Wang, Jie Zhu

Writing-review & editing: Zheng Wang, Jie Zhu, Lixian Ma

References

1. Torres J, Mehandru S, Colombel JF, Peyrin-Biroulet L. Crohn's disease. *Lancet* (London, England). 2017;389(10080):1741-55. [http://dx.doi.org/10.1016/s0140-6736\(16\)31711-1](http://dx.doi.org/10.1016/s0140-6736(16)31711-1). PMID: [27914655](#)
2. Molodecky NA, Soon IS, Rabi DM, Ghali WA, Ferris M, Chernoff G, et al. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. *Gastroenterology*. 2012;142(1):46-54.e42; quiz e30. <http://dx.doi.org/10.1053/j.gastro.2011.10.001>. PMID: [22001864](#)
3. Moller FT, Andersen V, Wohlfahrt J, Jess T. Familial risk of inflammatory bowel disease: a population-based cohort study 1977-2011. *The American journal of gastroenterology*. 2015;110(4):564-71. <http://dx.doi.org/10.1038/ajg.2015.50>. PMID: [25803400](#)
4. Baumgart DC, Carding SR. Inflammatory bowel disease: cause and immunobiology. *Lancet* (London, England). 2007;369(9573):1627-40. [http://dx.doi.org/10.1016/s0140-6736\(07\)60750-8](http://dx.doi.org/10.1016/s0140-6736(07)60750-8). PMID: [17499605](#)
5. Zhang YZ, Li YY. Inflammatory bowel disease: pathogenesis. *World journal of gastroenterology*. 2014;20(1):91-9. <http://dx.doi.org/10.3748/wjg.v20.i1.91>. PMID: [24415861](#)
6. Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*. 2001;411(6837):599-603. <http://dx.doi.org/10.1038/35079107>. PMID: [11385576](#)
7. Fritz T, Niederreiter L, Adolph T, Blumberg RS, Kaser A. Crohn's disease: NOD2, autophagy and ER stress converge. *Gut*. 2011;60(11):1580-8. <http://dx.doi.org/10.1136/gut.2009.206466>. PMID: [21252204](#)

- 479 8. Massey DC, Parkes M. Genome-wide association scanning highlights two autophagy genes,
480 ATG16L1 and IRGM, as being significantly associated with Crohn's disease. Autophagy.
481 2007;3(6):649-51. PMID: [17921695](#)
- 482 9. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, et al. Association analyses
483 identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across
484 populations. Nature genetics. 2015;47(9):979-86. <http://dx.doi.org/10.1038/ng.3359>. PMID:
485 [26192919](#)
- 486 10. Rung J, Brazma A. Reuse of public genome-wide gene expression data. Nature reviews Genetics.
487 2013;14(2):89-99. <http://dx.doi.org/10.1038/nrg3394>. PMID: [23269463](#)
- 488 11. Kolde R, Laur S, Adler P, Vilo J. Robust rank aggregation for gene list integration and
489 meta-analysis. Bioinformatics (Oxford, England). 2012;28(4):573-80.
490 <http://dx.doi.org/10.1093/bioinformatics/btr709>. PMID: [22247279](#)
- 491 12. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC
492 bioinformatics. 2008;9:559. <http://dx.doi.org/10.1186/1471-2105-9-559>. PMID: [19114008](#)
- 493 13. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using
494 gene-expression signatures to connect small molecules, genes, and disease. Science (New York, NY).
495 2006;313(5795):1929-35. <http://dx.doi.org/10.1126/science.1132939>. PMID: [17008526](#)
- 496 14. Lamb J. The Connectivity Map: a new tool for biomedical research. Nature reviews Cancer.
497 2007;7(1):54-60. <http://dx.doi.org/10.1038/nrc2044>. PMID: [17186018](#)
- 498 15. Gheeya J, Johansson P, Chen QR, Dexheimer T, Metaferia B, Song YK, et al. Expression profiling
499 identifies epoxy anthraquinone derivative as a DNA topoisomerase inhibitor. Cancer letters.
500 2010;293(1):124-31. <http://dx.doi.org/10.1016/j.canlet.2010.01.004>. PMID: [20133050](#)

- 501 16. Song Y, Liu J, Huang S, Zhang L. Analysis of differentially expressed genes in placental tissues of
502 preeclampsia patients using microarray combined with the Connectivity Map database. Placenta.
503 2013;34(12):1190-5. <http://dx.doi.org/10.1016/j.placenta.2013.09.013>. PMID: [24125805](#)
- 504 17. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open
505 software development for computational biology and bioinformatics. Genome biology.
506 2004;5(10):R80. <http://dx.doi.org/10.1186/gb-2004-5-10-r80>. PMID: [15461798](#)
- 507 18. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high
508 density oligonucleotide array data based on variance and bias. Bioinformatics (Oxford, England).
509 2003;19(2):185-93. PMID: [12538238](#)
- 510 19. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in
511 microarray experiments. Statistical applications in genetics and molecular biology. 2004;3:Article3.
512 <http://dx.doi.org/10.2202/1544-6115.1027>. PMID: [16646809](#)
- 513 20. Tripathi S, Pohl MO, Zhou Y, Rodriguez-Frandsen A, Wang G, Stein DA, et al. Meta- and
514 Orthogonal Integration of Influenza "OMICS" Data Defines a Role for UBR4 in Virus Budding. Cell host
515 & microbe. 2015;18(6):723-35. <http://dx.doi.org/10.1016/j.chom.2015.11.002>. PMID: [26651948](#)
- 516 21. Galamb O, Sipos F, Solymosi N, Spisak S, Krenacs T, Toth K, et al. Diagnostic mRNA expression
517 patterns of inflamed, benign, and malignant colorectal biopsy specimen and their correlation with
518 peripheral blood results. Cancer epidemiology, biomarkers & prevention : a publication of the
519 American Association for Cancer Research, cosponsored by the American Society of Preventive
520 Oncology. 2008;17(10):2835-45. <http://dx.doi.org/10.1158/1055-9965.epi-08-0231>. PMID: [18843029](#)
- 521 22. Arijis I, De Hertogh G, Lemaire K, Quintens R, Van Lommel L, Van Steen K, et al. Mucosal gene
522 expression of antimicrobial peptides in inflammatory bowel disease before and after first infliximab

523 treatment. PloS one. 2009;4(11):e7984. <http://dx.doi.org/10.1371/journal.pone.0007984>. PMID:

524 [19956723](https://pubmed.ncbi.nlm.nih.gov/19956723/)

525 23. Montero-Melendez T, Llor X, Garcia-Planella E, Perretti M, Suarez A. Identification of novel

526 predictor classifiers for inflammatory bowel disease by gene expression profiling. PloS one.

527 2013;8(10):e76235. <http://dx.doi.org/10.1371/journal.pone.0076235>. PMID: [24155895](https://pubmed.ncbi.nlm.nih.gov/24155895/)

528 24. Carey R, Jurickova I, Ballard E, Bonkowski E, Han X, Xu H, et al. Activation of an

529 IL-6:STAT3-dependent transcriptome in pediatric-onset inflammatory bowel disease. Inflammatory

530 bowel diseases. 2008;14(4):446-57. <http://dx.doi.org/10.1002/ibd.20342>. PMID: [18069684](https://pubmed.ncbi.nlm.nih.gov/18069684/)

531 25. Kugathasan S, Baldassano RN, Bradfield JP, Sleiman PM, Imielinski M, Guthery SL, et al. Loci on

532 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. Nature genetics.

533 2008;40(10):1211-5. <http://dx.doi.org/10.1038/ng.203>. PMID: [18758464](https://pubmed.ncbi.nlm.nih.gov/18758464/)

534 26. Vancamelbeke M, Vanuytsel T, Farre R, Verstockt S, Ferrante M, Van Assche G, et al. Genetic and

535 Transcriptomic Bases of Intestinal Epithelial Barrier Dysfunction in Inflammatory Bowel Disease.

536 Inflammatory bowel diseases. 2017;23(10):1718-29.

537 <http://dx.doi.org/10.1097/mib.0000000000001246>. PMID: [28885228](https://pubmed.ncbi.nlm.nih.gov/28885228/)

538 27. Wu F, Dassopoulos T, Cope L, Maitra A, Brant SR, Harris ML, et al. Genome-wide gene expression

539 differences in Crohn's disease and ulcerative colitis from endoscopic pinch biopsies: insights into

540 distinctive pathogenesis. Inflammatory bowel diseases. 2007;13(7):807-21.

541 <http://dx.doi.org/10.1002/ibd.20110>. PMID: [17262812](https://pubmed.ncbi.nlm.nih.gov/17262812/)

542 28. Leal RF, Planell N, Kajekar R, Lozano JJ, Ordas I, Dotti I, et al. Identification of inflammatory

543 mediators in patients with Crohn's disease unresponsive to anti-TNFalpha therapy. Gut.

544 2015;64(2):233-42. <http://dx.doi.org/10.1136/gutjnl-2013-306518>. PMID: [24700437](https://pubmed.ncbi.nlm.nih.gov/24700437/)

- 545 29. Perez-Torras S, Iglesias I, Llopis M, Lozano JJ, Antolin M, Guarner F, et al. Transportome Profiling
546 Identifies Profound Alterations in Crohn's Disease Partially Restored by Commensal Bacteria. Journal
547 of Crohn's & colitis. 2016;10(7):850-9. <http://dx.doi.org/10.1093/ecco-jcc/jjw042>. PMID: [26874350](#)
- 548 30. von Lampe B, Barthel B, Coupland SE, Riecken EO, Rosewicz S. Differential expression of matrix
549 metalloproteinases and their tissue inhibitors in colon mucosa of patients with inflammatory bowel
550 disease. Gut. 2000;47(1):63-73. PMID: [10861266](#)
- 551 31. Zeissig S, Burgel N, Gunzel D, Richter J, Mankertz J, Wahnschaffe U, et al. Changes in expression
552 and distribution of claudin 2, 5 and 8 lead to discontinuous tight junctions and barrier dysfunction in
553 active Crohn's disease. Gut. 2007;56(1):61-72. <http://dx.doi.org/10.1136/gut.2006.094375>. PMID:
554 [16822808](#)
- 555 32. Baugh MD, Perry MJ, Hollander AP, Davies DR, Cross SS, Lobo AJ, et al. Matrix metalloproteinase
556 levels are elevated in inflammatory bowel disease. Gastroenterology. 1999;117(4):814-22. PMID:
557 [10500063](#)
- 558 33. Fanjul-Fernandez M, Folgueras AR, Cabrera S, Lopez-Otin C. Matrix metalloproteinases:
559 evolution, gene regulation and functional analysis in mouse models. Biochimica et biophysica acta.
560 2010;1803(1):3-19. <http://dx.doi.org/10.1016/j.bbamcr.2009.07.004>. PMID: [19631700](#)
- 561 34. Ravi A, Garg P, Sitaraman SV. Matrix metalloproteinases in inflammatory bowel disease: boon or
562 a bane? Inflammatory bowel diseases. 2007;13(1):97-107. <http://dx.doi.org/10.1002/ibd.20011>.
563 PMID: [17206645](#)
- 564 35. Tsuchida C, Sakuramoto-Tsuchida S, Taked M, Itaya-Hironaka A, Yamauchi A, Misu M, et al.
565 Expression of REG family genes in human inflammatory bowel diseases and its regulation.

566 Biochemistry and biophysics reports. 2017;12:198-205.
567 <http://dx.doi.org/10.1016/j.bbrep.2017.10.003>. PMID: [29090282](#)

568 36. Lal-Nag M, Morin PJ. The claudins. Genome biology. 2009;10(8):235.
569 <http://dx.doi.org/10.1186/gb-2009-10-8-235>. PMID: [19706201](#)

570 37. Wang H, Chao K, Ng SC, Bai AH, Yu Q, Yu J, et al. Pro-inflammatory miR-223 mediates the
571 cross-talk between the IL23 pathway and the intestinal barrier in inflammatory bowel disease.
572 Genome biology. 2016;17:58. <http://dx.doi.org/10.1186/s13059-016-0901-8>. PMID: [27029486](#)

573 38. Brenna O, Bruland T, Furnes MW, Granlund A, Drozdov I, Emgard J, et al. The guanylate cyclase-C
574 signaling pathway is down-regulated in inflammatory bowel disease. Scandinavian journal of
575 gastroenterology. 2015;50(10):1241-52. <http://dx.doi.org/10.3109/00365521.2015.1038849>. PMID:
576 [25979109](#)

577 39. Amin AR, Attur MG, Thakker GD, Patel PD, Vyas PR, Patel RN, et al. A novel mechanism of action
578 of tetracyclines: effects on nitric oxide synthases. Proceedings of the National Academy of Sciences of
579 the United States of America. 1996;93(24):14014-9. PMID: [8943052](#)

580 40. Wang LX, Li Y, Chen GZ. Network-based co-expression analysis for exploring the potential
581 diagnostic biomarkers of metastatic melanoma. PloS one. 2018;13(1):e0190447.
582 <http://dx.doi.org/10.1371/journal.pone.0190447>. PMID: [29377892](#)

583 41. Yan S, Wang W, Gao G, Cheng M, Wang X, Wang Z, et al. Key genes and functional coexpression
584 modules involved in the pathogenesis of systemic lupus erythematosus. 2018;233(11):8815-25.
585 <http://dx.doi.org/10.1002/jcp.26795>. PMID: [29806703](#)

586 42. Xie D, Zhang Y, Qu H. Crucial genes of inflammatory bowel diseases explored by gene expression
587 profiling analysis. *Scandinavian journal of gastroenterology*. 2018;53(6):685-91.
588 <http://dx.doi.org/10.1080/00365521.2018.1461923>. PMID: [29909694](#)

589 43. Lin X, Li J, Zhao Q, Feng JR, Gao Q, Nie JY. WGCNA Reveals Key Roles of IL8 and MMP-9 in
590 Progression of Involvement Area in Colon of Patients with Ulcerative Colitis. *Current medical science*.
591 2018;38(2):252-8. <http://dx.doi.org/10.1007/s11596-018-1873-6>. PMID: [30074183](#)

592 44. Noisakran S, Carr DJ. Type I interferons and herpes simplex virus infection: a naked DNA
593 approach as a therapeutic option? *Immunologic research*. 2001;24(1):1-11.
594 <http://dx.doi.org/10.1385/ir:24:1:01>. PMID: [11485206](#)

595 45. Su C, Zhan G, Zheng C. Evasion of host antiviral innate immunity by HSV-1, an update. *Virology*
596 *journal*. 2016;13:38. <http://dx.doi.org/10.1186/s12985-016-0495-5>. PMID: [26952111](#)

597 46. Wagner J, Sim WH, Lee KJ, Kirkwood CD. Current knowledge and systematic review of viruses
598 associated with Crohn's disease. *Reviews in medical virology*. 2013;23(3):145-71.
599 <http://dx.doi.org/10.1002/rmv.1720>. PMID: [22674582](#)

600 47. Riley PA. Bacteriophages in autoimmune disease and other inflammatory conditions. *Medical*
601 *hypotheses*. 2004;62(4):493-8. <http://dx.doi.org/10.1016/j.mehy.2003.12.016>. PMID: [15050095](#)

602 48. Cadwell K, Patel KK, Maloney NS, Liu TC, Ng AC, Storer CE, et al. Virus-plus-susceptibility gene
603 interaction determines Crohn's disease gene Atg16L1 phenotypes in intestine. *Cell*.
604 2010;141(7):1135-45. <http://dx.doi.org/10.1016/j.cell.2010.05.009>. PMID: [20602997](#)

605 49. Ormeci AC, Akyuz F, Baran B, Soyer OM, Gokturk S, Onel M, et al. Steroid-refractory
606 inflammatory bowel disease is a risk factor for CMV infection. *European review for medical and*
607 *pharmacological sciences*. 2016;20(5):858-65. PMID: [27010142](#)

50. Specia S, Giusti I, Rieder F, Latella G. Cellular and molecular mechanisms of intestinal fibrosis.

World journal of gastroenterology. 2012;18(28):3635-61. <http://dx.doi.org/10.3748/wjg.v18.i28.3635>.

PMID: [22851857](https://pubmed.ncbi.nlm.nih.gov/22851857/)

51. Valatas V, Filidou E, Drygiannakis I, Kolios G. Stromal and immune cells in gut fibrosis: the

myofibroblast and the scarface. Annals of gastroenterology. 2017;30(4):393-404.

<http://dx.doi.org/10.20524/aog.2017.0146>. PMID: [28655975](https://pubmed.ncbi.nlm.nih.gov/28655975/)

Supporting information

S1 Table. Top 150 up-regulated and top 23 down-regulated genes in CD patients.

S2 Table. Six main functional co-expression modules and related genes involved in CD.

Fig1

Fig2A

Cluster Dendrogram

Fig2B

Network heatmap plot, all genes

bioRxiv preprint doi: <https://doi.org/10.1101/543835>; this version posted February 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Fig2C

Module-trait relationships

Fig3

Fig4