Genes & Immunity (2019) 20:131–142
https://doi.org/10.1038/s41435-018-0015-2

ARTICLE

Genetic variants and pathways implicated in a pediatric inflammatory bowel disease cohort

Kelly A. Shaw[1] · David J. Cutler[1] · David Okou[2] · Anne Dodd[2] · Bruce J. Aronow[3] · Yael Haberman[4] · Christine Stevens[5] · Thomas D. Walters[6] · Anne Griffiths[6] · Robert N. Baldassano[7] · Joshua D. Noe[8] · Jeffrey S. Hyams[9] · Wallace V. Crandall[10] · Barbara S. Kirschner[11] · Melvin B. Heyman[12] · Scott Snapper[13] · Stephen Guthery[14] · Marla C. Dubinsky[15] · Jason M. Shapiro[16] · Anthony R. Otley[17] · Mark Daly[5] · Lee A. Denson[4] · Subra Kugathasan[2] · Michael E. Zwick[1]

Received: 8 June 2017 / Revised: 28 August 2017 / Accepted: 11 September 2017 / Published online: 28 March 2018

Abstract

In the United States, approximately 5% of individuals with inflammatory bowel disease (IBD) are younger than 20 years old. Studies of pediatric cohorts can provide unique insights into genetic architecture of IBD, which includes Crohn's disease (CD) and ulcerative colitis (UC). Large genome-wide association studies have found more than 200 IBD-associated loci but explain a minority of disease variance for CD and UC. We sought to characterize the contribution of rare variants to disease development, comparing exome sequencing of 368 pediatric IBD patients to publicly available exome sequencing (dbGaP) and aggregate frequency data (ExAC). Using dbGaP data, we performed logistic regression for common variants and optimal unified association tests (SKAT-O) for rare, likely-deleterious variants. We further compared rare variants to ExAC counts with Fisher's exact tests. We did pathway enrichment analysis on the most significant genes from each comparison. Many variants overlapped with known IBD-associated genes (e.g. *NOD2*). Rare variants were enriched in CD-associated loci ($p = 0.009$) and showed suggestive enrichment in neutrophil function genes ($p = 0.05$). Pathway enrichment implicated immune-related pathways, especially cell killing and apoptosis. Variants in extracellular matrix genes also emerged as an important theme in our analysis.

Introduction

Crohn's disease (CD) and ulcerative colitis (UC) are the most common inflammatory bowel diseases (IBD) and are characterized by chronic remitting and relapsing gastrointestinal inflammation. In the United States, the prevalence of IBD for children (<20 years old) was estimated to be 92 cases per 100,000 in 2009, accounting for approximately 5% of prevalent cases [1]. Increasing prevalence [1] and rates of hospitalization [2] for pediatric IBD have been observed in the US, mirroring the trend of increasing IBD incidence in both pediatric [3, 4] and adult [5] populations worldwide. Diagnosed early in life, pediatric patients face years of medication, surveillance colonoscopy, and a high probability of surgery. Better understanding of disease etiology and progression in this group is therefore vital.

IBD is thought to have a strong genetic component, since family history of IBD is the greatest risk factor for disease at all ages. IBD patients with a family history of disease often present at a younger age [6–8], are more likely to experience extra-intestinal manifestations [6], have perforating disease, and require longer follow-up compared to patients without family history [6, 7], likely reflecting an increased genetic liability to disease. Genetic analyses of pediatric cohorts are therefore useful in exploring genetic architecture of IBD.

Large genome-wide association studies (GWAS) of IBD have found more than 200 common loci associated with disease [9, 10]. Pathway analysis of associated loci has

These authors contributed equally: Subra Kugathasan, Michael E. Zwick.

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41435-018-0015-2) contains supplementary material, which is available to authorized users.

✉ Michael E. Zwick
mzwick@emory.edu

Extended author information available on the last page of the article

Table 1 Clinical and demographic characteristics of samples with exome sequencing data used in analysis

	IBD cases of European ancestry	ARRA controls of European ancestry	Epi4k controls of European ancestry
Age at participation			
Range	0–17	18–84	Ages not provided, but controls were parents of children with epilepsy
Median	8	51	
Mean	7.3	52	
Gender			
Female	152 (41%)	118 (56%)	223 (53%)
Male	216 (59%)	91 (44%)	199 (47%)
Diagnosis			
CD	281 (76%)	—	—
UC	61 (17%)	—	—
IBD-other	26 (7%)	—	—

Dashes indicate not applicable

found an enrichment of immune system genes, especially those related to host response to microbes, and a great deal of overlap with other immune diseases [9]. Findings of studies of common variation in pediatric IBD cohorts generally echo findings in adult populations. One study of greater than 1000 pediatric-onset IBD cases and 1600 controls found slightly increased odds ratios for risk alleles also found in adult populations (including the well-known *NOD2*), and greater burden of these common variants was weakly correlated with earlier age of onset in CD [11].

A small proportion of disease liability has been explained by common variants in IBD—13.1% in CD and 8.2% in UC [9]—but the contribution of rare variants has not been assessed. This class of genetic variation is important because explosive growth of the human population in recent history has led to a corresponding excess of rare alleles [12], and most variants in protein-coding sequence are at low frequency [13–15]. The availability of public data sets allows us to compare whole-exome sequencing (WES) of a pediatric IBD cohort to other WES data [16] and to large databases containing population allele frequency information [15, 17]. We can further examine pathways implicated by genes annotated to these rare variants to gain greater understanding of IBD.

Results

Study participant characteristics

Relevant demographic and clinical characteristics are shown in Table 1 for the 368 cases with pediatric-onset IBD (<18 years of age at diagnosis) and 625 publicly available controls from the database of Genotypes and Phenotypes (dbGaP) whose data passed our quality control filters and principal components criteria (see Methods and Supplementary Fig. 1). The characteristics of the initial cohort of 517 pediatric-onset IBD cases (see Methods) are also available in Supplementary Table 1.

Common variants (MAF>0.05)

Using logistic regression to compare sites with minor allele frequency (MAF) > 0.05 between the 368 pediatric-onset IBD cases and 625 publicly available controls, we found no sites that reached genome-wide significance after genomic control ($p < 2E-06$, Figure 1 and Table 2). However, 14 out of the top 20 sites were within known CD- or IBD-associated loci (full list of loci from Jostins 2012 [9] and Liu 2015 [10] available as Supplementary Table 2). Nine variants were around the locus containing *CARD9*, a gene associated with both CD and UC (Supplementary Fig. 2), and three variants were near the locus containing CD-associated *NOD2*. Two protective variants also appeared at other CD loci in *ADAM30* and *NOTCH2*. Genes annotated to the top 20 sites that also appeared in our list of genes involved in neutrophil function (Supplementary Table 3) included *NOD2* and *CARD9*, which have key roles in anti-bacterial and anti-fungal functions of monocytes and macrophages.

Pathway enrichment

Many of the pathways we found in our ClueGO pathway enrichment analysis that were implicated by the top 200 most significant annotated genes were immune-related (Table 3 and Fig. 2). The largest network of significant gene ontology (GO) terms included regulation of production of molecular mediators of immune response, as well as regulation of cytokine and tumor necrosis factor production.

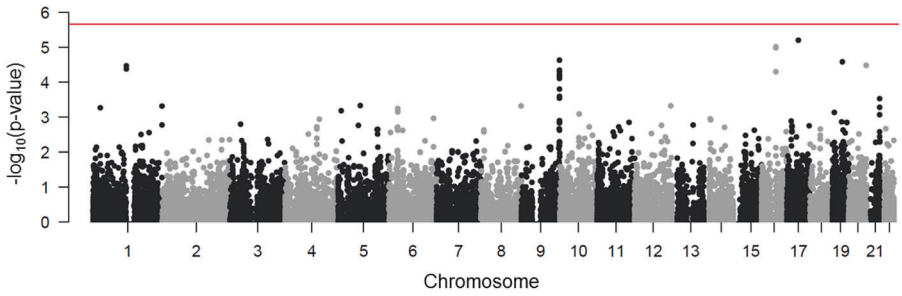

Fig. 1 Manhattan plot of *p*-values from logistic regression (with significant principal components and sex as covariates) comparing frequency of exome sequencing common variants in pediatric IBD cases to controls from dbGaP

Table 2 Top 20 most significant loci found in our common variant logistic regression

Chrom	Position	ID	Alt	Type	OR	Gene	*p*-value	Assoc. Diagnosis, Study	Neut. gene list
chr17	41227070	chr17_41227070	A	DEL	0.1111	KRTAP9-2	6.21E-06	—	—
chr16	50711288	rs2066843	T	SNP	1.57	NOD2	9.60E-06	CD, Jostins	Yes
chr16	50710713	rs2066842	T	SNP	1.569	NOD2	1.02E-05	CD, Jostins	Yes
chr9	136371953	rs10781499	A	SNP	1.524	CARD9	2.32E-05	IBD, Jostins	Yes
chr19	35488794	rs10410228	T	SNP	1.653	KRTDAP	2.59E-05	—	—
chr20	62346665	rs6143036	A	SNP	1.607	LAMA5	3.27E-05	—	—
chr1	119895261	rs2641348	G	SNP	0.4588	ADAM30	3.37E-05	CD, Jostins	—
chr1	119915381	rs6685892	T	SNP	0.4627	NOTCH2	4.14E-05	CD, Jostins	—
chr9	136372044	rs4077515	T	SNP	1.499	CARD9	4.49E-05	IBD, Jostins	Yes
chr16	50675812	rs6596	A	SNP	1.613	SNX20	4.97E-05	CD, Jostins	—
chr9	136395373	rs4266763	G	SNP	1.496	SNAPC4	5.02E-05	IBD, Jostins	—
chr9	136380752	rs3812570	C	SNP	1.49	SNAPC4	6.04E-05	IBD, Jostins	—
chr9	136380842	rs3812571	C	SNP	1.486	SNAPC4	6.77E-05	IBD, Jostins	—
chr9	136384721	rs10781510	A	SNP	1.484	SNAPC4	7.74E-05	IBD, Jostins	—
chr9	136404141	rs1051957	G	SNP	1.451	SDCCAG3	0.00016	IBD, Jostins	—
chr9	136477334	rs6560632	C	SNP	1.427	SEC16A	0.00025	IBD, Jostins	—
chr9	136432987	rs10781542	G	SNP	1.423	INPP5E	0.00029	IBD, Jostins	—
chr21	46246830	rs17183220	T	SNP	0.4371	MCM3AP-AS1	0.00029	—	—
chr5	78885600	rs1071598	T	SNP	1.558	ARSB	0.00046	—	—
chr12	128899303	chr12_128899303	G	DEL	0.4497	GLT1D1	0.00047	—	—

Dashes indicate not applicable or no.

Terms related to regulation of leukocyte-mediated immunity, cytotoxicity, and apoptosis were also significant. Other associated pathways related to the theme of cell killing included positive regulation of apoptotic cell clearance and regulation of complement activation. Regulation of keratinocyte proliferation, Ras signal transduction, and muscle cell and neural crest cell development were also implicated.

Rare variants (MAF<0.05)

Optimal unified association test (SKAT-O) analysis of rare variants

Using the same IBD and dbGaP cohorts, we tested rare variants with combined annotation dependent depletion (CADD) scores [18] greater than 10 to see if any genes were significantly enriched with these possibly pathogenic variants. The only genome-wide significant gene ($p < 2E-05$) was the well-known *NOD2* (Table 4A). When we tested enrichment of variants in loci associated with IBD, the only significant list was the Crohn's-disease-associated loci ($p = 0.009$, Table 4B). We also found a suggestive relationship between case status and rare variants in 144 genes that have been implicated in neutrophil function ($p = 0.05$, Table 4C).

We re-ran the SKAT-O analysis, adding common variants with CADD scores >10 to our list of rare variants. Including these common variants did not greatly impact the significance of genes associated with case status, likely because there were relatively few variants above the CADD score cutoff at 5% frequency or greater. However, including common variants strengthened the enrichment of variants in CD genes ($p = 0.004$; Supplementary Table 4A) and neutrophil function genes ($p = 0.03$; Supplementary Table 4B).

Table 3 Significantly enriched pathways in the top 200 most significant genes in our common variant (dbGaP) analysis

GO ID	GO term	% pathway covered	Corrected *p*-value	Associated genes found
GO:2000427	Positive regulation of apoptotic cell clearance	33	9E-05	[C2, C3, CCL2]
GO:0001910	Regulation of leukocyte-mediated cytotoxicity	9.4	2E-04	[CCL2, HLA-A, LILRB1, RASGRP1, SERPINB4]
GO:0002699	Positive regulation of immune effector process	4.5	5E-04	[C3, CCL2, GPI, HLA-A, IL2, LILRB1, NOD2, RASGRP1]
GO:0002703	Regulation of leukocyte-mediated immunity	4.4	0.001	[C3, HLA-A, IL2, LILRB1, NOD2, RASGRP1, SERPINB4]
GO:0055001	Muscle cell development	4.3	0.002	[ANK2, FHOD3, GPX1, IGSF22, MYPN, RYR1, XK]
GO:0001578	Microtubule bundle formation	5.8	0.002	[CCDC40, DNAH5, MAP1B, RP1L1, SPAG16]
GO:0002705	Positive regulation of leukocyte-mediated immunity	5.5	0.003	[C3, HLA-A, IL2, NOD2, RASGRP1]
GO:0048747	Muscle fiber development	7.1	0.003	[GPX1, MYPN, RYR1, XK]
GO:0014032	Neural crest cell development	6.9	0.003	[ERBB4, JAG1, LAMA5, RET]
GO:0010927	Cellular component assembly involved in morphogenesis	5.2	0.003	[ANK2, DAG1, FHOD3, IGSF22, MYPN]
GO:0046487	Glyoxylate metabolic process	9.7	0.004	[AMT, LDHD, LIAS]
GO:0006081	Cellular aldehyde metabolic process	4.8	0.005	[AMT, GPI, H6PD, LDHD, LIAS]
GO:0032680	Regulation of tumor necrosis factor production	4.7	0.005	[CARD9, CCL2, LILRB1, NOD2, RASGRP1]
GO:0010837	Regulation of keratinocyte proliferation	8.8	0.005	[EPPK1, STXBP4, TGM1]
GO:0001912	Positive regulation of leukocyte-mediated cytotoxicity	8.6	0.006	[CCL2, HLA-A, RASGRP1]
GO:0030449	Regulation of complement activation	8.1	0.007	[C2, C3, CFB]
GO:0002700	Regulation of production of molecular mediator of immune response	4.1	0.009	[GPI, HLA-A, IL2, LILRB1, NOD2]
GO:0045214	Sarcomere organization	6.8	0.01	[FHOD3, IGSF22, MYPN]
GO:0072676	Lymphocyte migration	4.8	0.01	[AIRE, CCL2, GCSAML, RET]
GO:2000106	Regulation of leukocyte apoptotic process	4.3	0.02	[IL2, LILRB1, NOD2, TP53BP1]
GO:0032649	Regulation of interferon-gamma production	4.2	0.02	[HLA-A, IL2, LILRB1, RASGRP1]
GO:0046579	Positive regulation of Ras protein signal transduction	5.6	0.02	[ARRB1, NOTCH2, RASGRP1]
GO:0002718	Regulation of cytokine production involved in immune response	4.2	0.04	[HLA-A, LILRB1, NOD2]

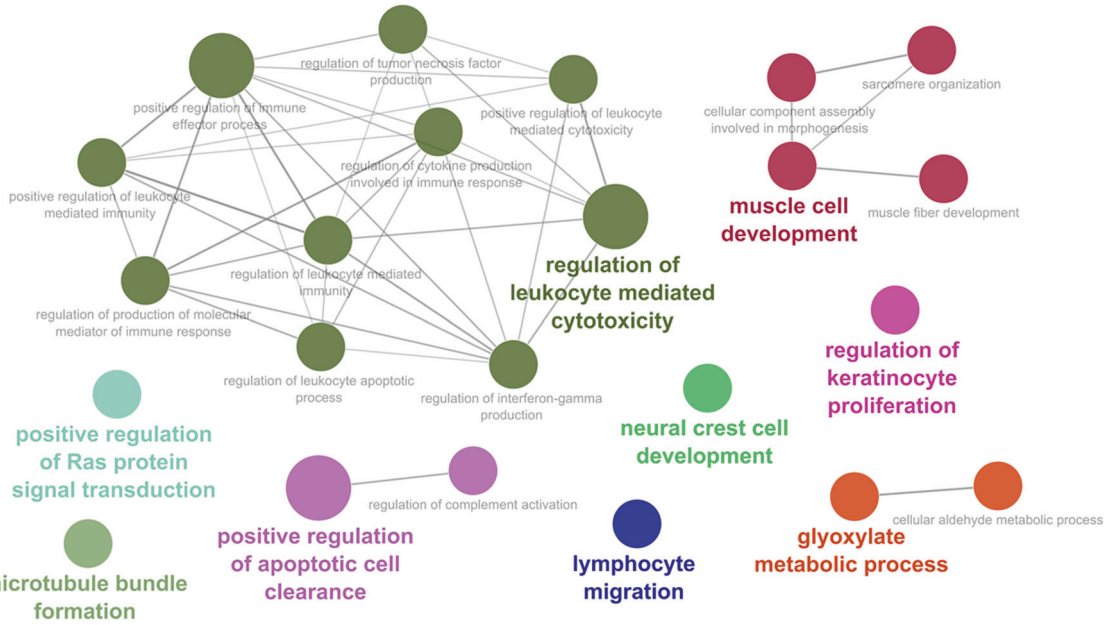

Fig. 2 Pathway enrichment of the genes annotated to the top 200 most significant common genes tested in our logistic regression

Exome Aggregation Consortium (ExAC) rare variant analysis

There was unsurprisingly a great deal of inflation when we performed Fisher's exact tests comparing rare variant counts between the 368 pediatric IBD patients and aggregate allele frequencies for Caucasian populations in the ExAC database (Supplementary Fig. 3). We therefore limited our analysis to sites that made it past the stringent QC in our dbGaP analysis, and further filtered out sites in ExAC that were most significantly different from our dbGaP controls (see Methods). As seen in Fig. 3, genome-wide inflation was no longer apparent after applying these criteria. Shown in Table 5, six variants were genome-wide significant ($p <$ 6E-07), with the most significant annotated to *NOD2*. Two other of the top 20 most significant variants were annotated to known IBD loci: one other in *NOD2* and one in *D2HGDH*. Of our list of neutrophil function genes, only *NOD2* was among the top 20 most significant rare variants.

Pathway enrichment

According to analysis in ClueGO, the top 200 most significant genes in our list of rare variants were enriched in a few pathways (Table 6 and Fig. 4). Immune-response-related hits included negative regulation of the JAK-STAT cascade, modulation by host of viral transcription, and modification by host of symbiont morphology and physiology. Genes were also enriched in pathways involving ion transmembrane transport and negative regulation of axon extension. ToppFun analysis also highlighted genes involved in response to bacterium, regulation of antigen

Table 4A Top 15 results from SKAT-O analysis of enrichment of rare, likely-pathogenic (CADD > 10) variants in genes with five or more variants

SetID	*p*-value	Number of variants included in gene
NOD2	8.4E-12	15
VWA2	0.0006	7
HAPLN3	0.0008	5
LMF1	0.002	5
SOS1	0.002	5
MAGI2	0.002	7
SRRM2	0.002	13
RGS12	0.003	10
SCAF4	0.003	5
STARD13	0.004	8
RHPN2	0.005	6
D2HGDH	0.005	6
G6PC2	0.005	6
NR4A1	0.005	5
EFEMP2	0.006	5

Table 4B SKAT-O analysis for enrichment of rare variants with CADD scores >10 in loci associated with Crohn's disease (CD), inflammatory bowel disease (IBD), or ulcerative colitis (UC)

SetID	*p*-value	Number of variants included in SetID
CD	0.009	522
IBD	0.9	1849
UC	0.7	445

Table 4C SKAT-O analysis for enrichment of rare, conserved variants in neutrophil function genes (NEUT)

SetID	*p*-value	Number of variants included in SetID
NEUT	0.05	413

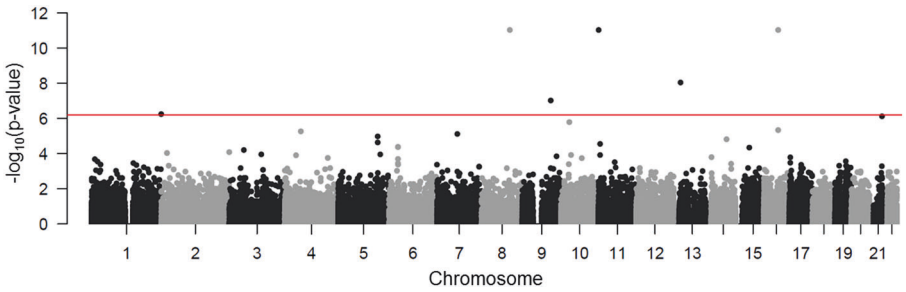

Fig. 3 Manhattan plot of *p*-values from comparing frequency of exome sequencing rare variants in pediatric IBD cases to ExAC after filtering out sites most significantly different between ExAC and our control data set

processing and presentation of peptide antigen, immune system development, and biological adhesion pathways (Supplementary Table 5).

Discussion

Our findings echo important aspects of previous genetic and pathway enrichment analyses. Crohn's-disease-associated loci had a strong showing in our results: two variants in *NOD2* were the most significant in our dbGaP common variant analysis, and one site was significant in our ExAC rare variant analysis. *NOD2* also emerged as significant in our gene-level SKAT-O analysis, and CD-associated genes as a group were also significant. This was not unexpected since the majority of our cohort were Crohn's patients. Of the top 20 most significant common variants, 9 were within a single 100 kb region around *CARD9* (Supplementary Fig. 2), a gene that has long been associated with IBD. This entire region looks equally associated with disease (OR ~1.5) in our cohort, reflecting that deep sequencing still cannot solve problems regarding fine mapping of causative variants without sufficient recombination.

We also found intriguing variants in genes not yet associated with IBD. *KRTAP9-2* and *KRTDAP*, two of our top five common variant findings, are involved in keratinocyte differentiation, a theme that also emerged in our common variant pathway analysis. Keratinocytes are the most abundant component of the epidermis, playing an important role in immunomodulation at the interface between the body and environment. Capable of producing cytokines, these cells have been linked to a different inflammatory disease, psoriasis [19, 20]. Additionally, one recent study found that the interplay of hair follicle development, colonization by commensal microbiota, and local chemokine production in skin was necessary to establish immune tolerance to commensal microbes [21]; dysfunction in the skin environment could potentially impact this process and have systemic immune repercussions. These suggestive findings require replication in future, larger studies of pediatric IBD.

LAMA5, another top hit in our common variant analysis, encodes a subunit of laminin. Laminins are extracellular matrix proteins which are a major component of the basement membrane, a matrix of tissue that separates the epithelium, mesothelium, and endothelium from underlying connective tissue. Because of the important role of laminins in the integrity of this layer, there could be a role for *LAMA5* in IBD pathogenesis. One study of transgenic mice overexpressing the *LAMA5* mouse homolog found an attenuated response to DSS-induced inflammation [22]. The two most significant genes in our SKAT-O rare variant analysis after *NOD2*, *VWA2* and *HAPLN3*, are also extracellular matrix components. In addition, the location and functions of the products of these genes are linked to integrins, which have emerged as important in large IBD GWAS [23]. And one recent, prospective study of more than 900 CD patients found that stricturing complications were associated with increased expression of extracellular matrix genes in ileal tissue at diagnosis [24]. Further studies are warranted to investigate the roles of these extracellular matrix proteins in disease etiology.

We were additionally interested in testing enrichment of rare variants in neutrophil function genes because children with inherited disorders of these classes of immune cells exhibit chronic intestinal inflammation similar to CD during the first decade of life [25, 26]. Similarly, loss of function in monocyte and/or macrophage antimicrobial pathways could be one mechanism of pediatric CD pathogenesis. Though we did not find a significant association, we did find a suggestive relationship in SKAT-O between rare, likely-deleterious variants in genes involved in neutrophil function and case status ($p = 0.05$). And when likely-deleterious common variants were also included, this association was significant ($p = 0.03$). Positive regulation of leukocyte-mediated immunity was also one of the most significant pathways in our common variant analysis, supporting

Table 5 Top 20 most significant sites in our rare variant Fisher's exact tests

Chrom	Position	ID	Alt	Type	OR	Gene	*p*-value	Assoc Diagnosis, Study	Neut. gene list
chr16	50729867	rs796661546	GC	INS	4.42151	NOD2	9.19E-12	CD, Jostins	Yes
chr8	100712766	chr8_100712766	CA	INS	33.764	PABPC1	9.19E-12	—	—
chr11	294540	chr11_294540	GC	INS	122.892	ATHL1	9.19E-12	—	—
chr13	24447181	chr13_24447181	T	DEL	700.496	PARP4	8.90E-09	—	—
chr9	101390469	chr9_101390469	GTA	INS	172.978	MRPL50	9.39E-08	—	—
chr1	248273809	chr1_248273809	C	DEL	279.84	OR2T33	5.58E-07	—	—
chr21	44573789	rs9977039	G	SNP	5.75281	TSPEAR	7.43E-07	—	—
chr10	29462394	chr10_29462394	AT	INS	Inf	SVIL-AS1	1.59E-06	—	—
chr16	50722629	rs2066845	C	MULTIALLELIC	3.4408	NOD2	4.52E-06	CD, Jostins	Yes
chr4	56964497	rs17087307	C	SNP	0.34222	NOA1	5.34E-06	—	—
chr7	72713798	rs146095374	A	SNP	0.25988	TYW1B	7.50E-06	—	—
chr5	140822334	rs61730632	A	SNP	2.79249	PCDHA1	1.03E-05	—	—
chr14	73953419	rs778985097	AT	INS	10.2459	COQ6	1.50E-05	—	—
chr5	140875534	rs114654172	G	SNP	2.70029	PCDHA1	2.28E-05	—	—
chr11	5544676	rs7934354	G	SNP	0.17691	OR52H1	2.76E-05	—	—
chr6	31960262	rs11541400	G	SNP	5.1923	SKIV2L	4.06E-05	—	—
chr6	31728544	rs139006870	A	SNP	5.17682	DDAH2	4.17E-05	—	—
chr15	49588022	chr15_49588022	CT	INS	Inf	FAM227B	4.47E-05	—	—
chr3	51995472	rs371570896	A	SNP	77.1658	RPL29	6.15E-05	—	—
chr2	241767780	rs143940595	A	SNP	0	D2HGDH	8.10E-05	CD, Liu	—

Dashes indicate not applicable or no

further study into the role of phagocyte function and dysfunction in IBD.

Another important component of the immune system from our pathway analysis was complement; mutations in C2, C3, and CFB were among the top 200 most significant common variants associated with disease in our cohort. Though research into the role of complement has been somewhat lacking, evidence is growing for its potential relevance in disease pathophysiology (reviewed in [27]). A closely related theme, apoptosis, also appeared in several other significant pathways.

Ras signaling was another pathway of interest from our common variant analysis, and *SOS1*, one of the top hits in our rare variant SKAT-O analysis, is also a guanine nucleotide exchange factor for RAS proteins. In fact, this pathway was previously implicated by a large study drawing from over 30,000 cases and 50,000 controls in contributing to IBD etiology as part of growth factor signaling [28]. Because growth factor deficiencies have been found in patients with IBD, there has been substantial interest in their use as a potential therapeutic agent (reviewed in [29]). Other current targets of therapy that emerged in our analysis include interferon-gamma, a pro-inflammatory cytokine involved in intestinal homeostasis and linked to regulation of IL-23 [30], another cytokine associated not only with IBD but other inflammatory diseases. In our rare variant analysis, we found negative regulation of the JAK-STAT cascade, another important inflammatory pathway targeted by recent therapies [31], which underscores the importance of immune cell response to cytokine signaling in disease.

The primary limitation of this study is the lack of in-house controls for comparison to our cases. However, we performed stringent QC of our data to filter differences between data sets. We used the same processing pipeline for dbGaP as we used for our case data, and filtered to an ancestrally similar population. However, systematic calling differences between our pipeline and ExAC, such as calling or filtering of indels, could still be leading to inflation of *p*-values and odds ratios in our rare variant analysis.

We combined CD and UC to leverage the maximum sample size possible to gain further insight into the shared genetic architecture of IBD. However, CD-related variants were enriched in our results, likely because of our CD-majority cohort and the large effect size of associated loci including *NOD2*. We still found variants in *HLA* genes, which are most strongly linked to UC, in our results, but these sites did not reach genome-wide significance in our cohort. For example, *HLA-A* and *HLA-C* were among the top 200 most significant genes in our logistic case/dbGaP control regression, and were therefore used in ClueGO analysis.

While large genome-wide association studies have been performed in IBD, our study is the first to specifically investigate the contribution of rare, likely-damaging

Table 6 Significantly enriched pathways using the list of the top 200 most significant genes in our ExAC rare variant analysis

GO ID	GO Term	% Pathway covered	Corrected *p*-value	Associated genes found
GO:0043921	Modulation by host of viral transcription	11	0.006	[HMGA2, POU2F3, PSG1]
GO:0030517	Negative regulation of axon extension	12	0.008	[BCL11A, RTN4R, SEMA5A]
GO:0051851	Modification by host of symbiont morphology or physiology	5.8	0.01	[HMGA2, POU2F3, PSG1, SMC3]
GO:0000288	Nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay	5.1	0.01	[EIF4A1, EIF4B, PABPC1, SKIV2L]
GO:0015698	Inorganic anion transport	4.1	0.01	[ABCB11, ANKH, CLCN6, CLCNKB, SLC12A6, SLC26A2, SLC5A5]
GO:0046426	Negative regulation of JAK-STAT cascade	5.7	0.02	[HMGA2, RTN4R, RTN4RL2]
GO:1902476	Chloride transmembrane transport	4.3	0.02	[CLCN6, CLCNKB, SLC12A6, SLC26A2]

variants in pediatric-onset disease. Our findings provide further targets for exploring disease etiology—both at the gene and pathway level. Better understanding of the genetic architecture of IBD can hopefully improve disease prediction and treatment.

Subjects and methods

Ethical approval and recruitment of study participants

Subjects for WES were selected from patients enrolled in the Crohn's and Colitis Foundation (CCFA) sponsored RISK cohort study and the NIH sponsored Emory African-American gene discovery study, for whom DNA had already been collected. RISK is the largest pediatric CD inception cohort in the world, with 1813 subjects younger than 18 years old with suspected IBD enrolled at 28 North American sites, including Emory University, from November 2008 to June 2012 (ClinicalTrials.gov Identifier: NCT00790543). All patients underwent baseline colonoscopy and histological confirmation of chronic active colitis/ileitis prior to diagnosis and treatment. Once standard and published guidelines were met, patients were diagnosed with CD, UC or IBD-undetermined (IBD-U). A consistent diagnosis of IBD was required during the one-year follow-up for inclusion into this study. At enrollment and during ongoing prospective follow-up, clinical and laboratory data were obtained for each enrolled patient and submitted to a centralized data management center. All patients were managed according to the dictates of their physicians, not by standardized protocols. The patient-based studies were approved by the Institutional Review Boards at each of the RISK sites. Consent was obtained from parents and adult subjects and assent from pediatric subjects age 11 and above.

Emory case sample collection, processing and exome sequencing

Genomic DNA was extracted from whole blood for a total of 567 pediatric IBD samples, of which 553 (97.5%) passed DNA QC. Library preparation and sequencing of the samples were performed at Broad Institute's Genomics Platform, Cambridge, USA. The libraries were prepared according to the manufacturer's instructions using 1 μg of input DNA per sample. DNA was subjected to whole-exome capture with the SureSelect Human All Exon 50-Mb Kit (Agilent Technologies) following the standard protocols. Library validation was done with the KAPA Library Quantification Kit (KAPA Biosystems) and the whole-

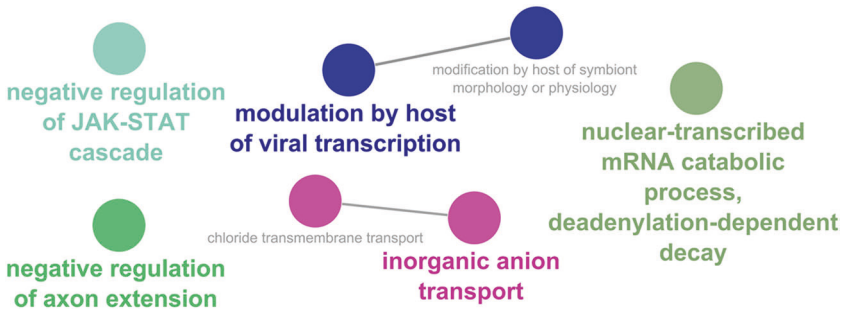

Fig. 4 Pathway enrichment of the genes annotated to the top 200 most significant rare variants tested in our rare variant analysis

exome capture libraries were then sequenced on the Illumina HiSeq platform according to standard protocols.

Publicly available data sets

Database of genotypes and phenotypes (dbGaP) [16] data

We identified and downloaded control data from the Epi4K (accession phs000653.v2.p1) and ARRA (accession phs000298.v3.p2) studies. SRA files were converted to fastq format using NCBI's SRA Toolkit [32].

ExAC (http://exac.broadinstitute.org/) [15, 17] data (version 0.3.1)

For this publicly available data set containing information on 60,706 individuals, we used liftOver to map all sites to hg38 for comparison with our data. We summed minor and total allele counts for the American, Finnish, and non-Finnish European groups and required a site to be typed in >90% of total chromosomes for these groups (at least 76,438 out of 84,930 chromosomes) for inclusion.

dbGaP (raw whole-exome sequencing) analysis

We mapped Emory and dbGaP exome sequencing fastq files to hg38 using PEMapper and called variants using PECaller [33]. We then used SeqAnt [34] version 2.0 [35] (Beta 3, https://seqant.genetics.emory.edu/) to get rsID numbers for plink and other annotation information for later analysis.

All following variant quality control (QC) was performed in PLINK 1.9 [36–38]. Starting with 866,411 variants in 1035 controls and 541 cases diagnosed with IBD before age 18, we filtered samples and variants using increasingly stringent completeness criteria until information for all remaining variants and samples was 99% complete. For each study individually (IBD, ARRA, Epi4k), we removed sites that were Bonferroni significant in a Hardy–Weinberg equilibrium test. We then performed a sex check of samples. Cases were removed if their sex was discordant with record review ($N = 9$); other mislabeled sexes were corrected. We checked sample relatedness and removed 8 controls and 10 cases who were second degree or more closely related to another study participant. Supplementary Table 1 shows characteristics for the 517 remaining IBD patients who passed this first round of quality control. We combined CD and UC patients because of shared genetic architecture of these diseases and relatively small sample size of either group alone.

To adjust for population stratification in our sample we used 10,913 common (minor allele frequency, a.k.a. MAF > 0.05) SNPs to calculate principal components (PCs) using EIGENSTRAT [39] and anchoring with HapMap controls as described by Anderson et al. [40] (Supplementary Fig. 1A). We removed outliers (those with values greater or less than 3 standard deviations away from the mean) for any of the top seven principal components (those which appeared meaningful with eigenvalues >2), recalculated principal components, and repeated outlier filtering with four meaningful PCs, leaving us with a final data set of 625 controls and 368 cases (Supplementary Fig. 1B; Table 1 shows basic characteristics for these participants). PCs were recalculated again without HapMap samples (Supplementary Fig. 1C) and the four principal components significant by Tracy-Widom tests were used as covariates in regressions.

As an additional filter, we removed variants that were most significantly different (top 2.5%) in Fisher's exact tests comparing our dbGaP controls to ExAC.

Common variant analysis

We performed logistic regression for sites with MAF > 0.05 in plink with case/control status as outcome, genotype as predictor of interest, and sex and PCs as covariates. *p*-Values were corrected with genomic control.

SKAT-O analysis

We used the SKAT-O method within the SKAT package [41] in R [42] to analyze genes annotated to sites with MAF < 0.05 and evidence of pathogenicity with CADD score >10. SKAT-O is an approach that optimizes association tests by unifying burden and sequence kernel association

approaches [43]. We tested for association of genes with case/control status for any gene with five or more rare variants. We also lifted over loci associated with IBD from Jostins et al. 2012 [9] and Liu et al. 2015 [10] to hg38, yielding 201 loci, and tested for enrichment of rare variants 250 kb upstream or downstream of CD, UC, or IBD loci as groups (Supplementary Table 2).

We also wanted to test whether variants were enriched in neutrophil function genes because strong ileal activation of the immune response including a strong signature for blood CD11b+Ly6-G+neutrophils (GSM854306, $p < 6.5E-50$) was found using clinical and RNA-Seq data from the CCFA RISK prospective cohort [44]. We next used the GSM854306 from immgen atlas (GSE15907) to retrieve all 409 blood CD11b+Ly6-G+neutrophil genes and combined this with a manually curated, literature-based list of 74 human neutrophil-related genes, including those known to cause CGD and GSD1b. We implemented these two gene lists in ToppCluster [45], cross-validating their association with neutrophil-related genes and pathways based on other annotations of critical neutrophil functions including priming, chemotaxis, adhesion, phagocytosis, oxidative burst, degranulation, microbial killing, and survival (GO, Mouse phenotypes, Diseases). Using this filtering we were able to decrease the original total of 463 neutrophil genes to 144 genes that are associated with CD and known to regulate key neutrophil functions (Supplementary Table 3).

ExAC (aggregate allele count) analysis

Rare variant analysis

Using the same set of variants as in the dbGaP analysis (with sites most significantly different between dbGaP and ExAC filtered out), we used Fisher's exact tests to compare rare variant sites (MAF < 0.05) between our IBD cases and ExAC. Genomic control was used to correct *p*-values.

Pathway enrichment analysis

To test for pathway enrichment, we used the ClueGO plugin version 2.3.3 for Cytoscape version 3.4.0. We performed right-sided hypergeometric tests for enrichment of level 3 to 8 biological process GO terms (using the Human GO database from 25 January 2017) with Benjamini–Hochberg *p*-value correction for multiple tests. GO Term Fusion was used to reduce pathway redundancy. For common and rare variants, the top 200 most significant genes were used to interrogate pathway enrichment in our sample. This threshold was picked so that ClueGO input did not have duplicate genes and was consistent across common and rare variant comparisons. All genes in the common variant analysis had *p*-values ≤0.01, while those in the rare variant analysis had $p \leq 0.002$.

We also used ToppFun, from the ToppGene Suite of bioinformatics tools, to perform functional enrichment analysis. While we only used biological process terms with ClueGO, ToppFun pulls annotation information from GO, human and mouse phenotype data, gene expression, protein interaction and pathway databases [46].

Data availability

Raw sequencing data for individuals with inflammatory bowel disease included in this study are publicly available on dbGaP. Study accession: phs001076.v1.p1, URL: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001076.v1.p1.

Acknowledgements We would like to thank our study participants as well as the providers and facilitators of publicly available data used in this project: the Exome Aggregation Consortium; the NCBI database of genotypes and phenotypes; the Epi4K Gene Discovery in Epilepsy study (NINDS U01-NS077303) and the Epilepsy Genome/Phenome Project (EPGP—NINDS U01-NS053998); and the ARRA Autism Sequencing Collaborative (an ARRA funded research initiative—R01-MH089208, R01-MH089175, R01-MH089025, R01-MH089004, and R01-MH089482).

Funding This work was supported by NIH R01 DK098231 and the Crohn's and Colitis Foundation. KAS received support from BWF training grant ID#1008188 and NIH NRSA F31 DK107229.

Compliance with ethical standards

Conflict of interest: The authors declare they have no conflict of interest.

References

1. Kappelman MD, Moore KR, Allen JK, Cook SF. Recent trends in the prevalence of Crohn's disease and ulcerative colitis in a commercially insured US population. Dig Dis Sci. 2013;58:519–25.

2. Sandberg KC, Davis MM, Gebremariam A, Adler J. Increasing hospitalizations in inflammatory bowel disease among children in the United States, 1988-2011. Inflamm Bowel Dis. 2014;20:1754–60.
3. Benchimol EI, Fortinsky KJ, Gozdyra P, Van den Heuvel M, Van Limbergen J, Griffiths AM. Epidemiology of pediatric inflammatory bowel disease: a systematic review of international trends. Inflamm Bowel Dis. 2011;17:423–39.
4. Benchimol EI, Bernstein CN, Bitton A, Carroll MW, Singh H, Otley AR et al. Trends in epidemiology of pediatric inflammatory bowel disease in Canada: distributed network analysis of multiple population-based provincial health administrative databases. Am J Gastroenterol. 2017. https://doi.org/10.1038/ajg.2017.97.
5. 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;142:46–54. e42quiz e30
6. Andreu M, Márquez L, Domènech E, Gisbert JP, García V, Marín-Jiménez I, et al. Disease severity in familial cases of IBD. J Crohns Colitis. 2014;8:234–9.
7. Carbonnel F, Macaigne G, Beaugerie L, Gendre JP, Cosnes J. Crohn's disease severity in familial and sporadic cases. Gut. 1999;44:91–5.
8. Henriksen M, Jahnsen J, Lygren I, Vatn MH, Moum B, IBSEN Study Group. Are there any differences in phenotype or disease course between familial and sporadic cases of inflammatory bowel disease? Results of a population-based follow-up study. Am J Gastroenterol. 2007;102:1955–63.
9. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491:119–24.
10. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet. 2015;47:979–86.
11. Cutler DJ, Zwick ME, Okou DT, Prahalad S, Walters T, Guthery SL, et al. Dissecting allele architecture of early onset IBD using high-density genotyping. PLoS ONE. 2015;10:e0128074.
12. Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science. 2012;336:740–3.
13. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012;337:64–9.
14. Marth GT, Yu F, Indap AR, Garimella K, Gravel S, Leong WF, et al. The functional spectrum of low-frequency coding variation. Genome Biol. 2011;12:R84.
15. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.
16. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet. 2007;39:1181–6.
17. Exome Aggregation Consortium (ExAC). Cambridge, MA 2015. http://exac.broadinstitute.org.
18. Johansen C, Rittig AH, Mose M, Bertelsen T, Weimar I, Nielsen J, et al. STAT2 is involved in the pathogenesis of psoriasis by promoting CXCL11 and CCL5 production by keratinocytes. PLoS ONE. 2017;12:e0176994.
19. Lowes MA, Suárez-Fariñas M, Krueger JG. Immunology of psoriasis. Annu Rev Immunol. 2014;32:227–55.
20. Scharschmidt TC, Vasquez KS, Pauli ML, Leitner EG, Chu K, Truong H-A, et al. Commensal microbes and hair follicle morphogenesis coordinately drive Treg migration into neonatal skin. Cell Host Microbe. 2017;21:467.e5
21. Spenlé C, Lefebvre O, Lacroute J, Méchine-Neuville A, Barreau F, Blottière HM, et al. The laminin response in inflammatory bowel disease: protection or malignancy? PloS ONE. 2014;9:e111336.
22. de Lange KM, Moutsianas L, Lee JC, Lamb CA, Luo Y, Kennedy NA, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat Genet. 2017;49:256–61.
23. Kugathasan S, Denson LA, Walters TD, Kim M-O, Marigorta UM, Schirmer M, et al. Prediction of complicated disease course for children newly diagnosed with Crohn's disease: a multicentre inception cohort study. Lancet Lond Engl. 2017;389:1710–8.
24. Rieber N, Hector A, Kuijpers T, Roos D, Hartl D. Current concepts of hyperinflammation in chronic granulomatous disease. Clin Dev Immunol. 2012;2012:252460.
25. Yu JE, De Ravin SS, Uzel G, Landers C, Targan S, Malech HL, et al. High levels of Crohn's disease-associated anti-microbial antibodies are present and independent of colitis in chronic granulomatous disease. Clin Immunol. 2011;138:14–22.
26. Jain U, Otley AR, Van Limbergen J, Stadnyk AW. The complement system in inflammatory bowel disease. Inflamm Bowel Dis. 2014;20:1628–37.
27. Li J, Wei Z, Chang X, Cardinale CJ, Kim CE, Baldassano RN, et al. Pathway-based genome-wide association studies reveal the association between growth factor activity and inflammatory bowel disease. Inflamm Bowel Dis. 2016;22:1540–51.
28. Krishnan K, Arnone B, Buchman A. Intestinal growth factors: potential use in the treatment of inflammatory bowel disease and their role in mucosal healing. Inflamm Bowel Dis. 2011;17:410–22.
29. Sheikh SZ, Matsuoka K, Kobayashi T, Li F, Rubinas T, Plevy SE. Cutting edge: IFN-gamma is a negative regulator of IL-23 in murine macrophages and experimental colitis. J Immunol (Baltim MD 1950). 2010;184:4069–73.
30. Pedersen J, Coskun M, Soendergaard C, Salem M, Nielsen OH. Inflammatory pathways of importance for management of inflammatory bowel disease. World J Gastroenterol. 2014;20:64–77.
31. NCBI SRA Toolkit. https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software.
32. Johnston HR, Chopra P, Wingo TS, Patel V, International Consortium on Brain and Behavior in 22q11.2 Deletion Syndrome, Epstein MP et al. PEMapper and PECaller provide a simplified approach to whole-genome sequencing. Proc Natl Acad Sci USA. 2017. https://doi.org/10.1073/pnas.1618065114.
33. Shetty AC, Athri P, Mondal K, Horner VL, Steinberg KM, Patel V, et al. SeqAnt: a web service to rapidly identify and annotate DNA sequence variations. BMC Bioinformatics. 2010;11:471.
34. Kotlar AV, Trevino CE, Zwick ME, Cutler DJ, Wingo TS. SeqAnt 2.0: Whole-genome annotation and natural-language searching in the cloud. bioRxiv 2017: 146514.
35. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4:7.
36. Purcell S, Chang C. PLINK 1.9. https://www.cog-genomics.org/plink2.
37. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
38. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38:904–9.

39. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nat Protoc. 2010;5:1564–73.
40. Lee S, Miropolsky L, Wu M. SKAT: SNP-Set (Sequence) Kernel Association Test. 2016 https://CRAN.R-project.org/package=SKAT.
41. R Core team. R: A language and environment for statistical computing. R Found Stat Comput Vienna Austria, 2015. http://www.R-project.org/.
42. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46:310–5.
43. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet. 2012;91:224–37.
44. Haberman Y, Tickle TL, Dexheimer PJ, Kim M-O, Tang D, Karns R, et al. Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. J Clin Invest. 2014;124:3617–33.
45. Kaimal V, Bardes EE, Tabar SC, Jegga AG, Aronow BJ. ToppCluster: a multiple gene list feature analyzer for comparative enrichment clustering and network-based dissection of biological systems. Nucleic Acids Res. 2010;38:W96–102.
46. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009;37:W305–311.

Affiliations

Kelly A. Shaw[1] · David J. Cutler[1] · David Okou[2] · Anne Dodd[2] · Bruce J. Aronow[3] · Yael Haberman[4] · Christine Stevens[5] · Thomas D. Walters[6] · Anne Griffiths[6] · Robert N. Baldassano[7] · Joshua D. Noe[8] · Jeffrey S. Hyams[9] · Wallace V. Crandall[10] · Barbara S. Kirschner[11] · Melvin B. Heyman[12] · Scott Snapper[13] · Stephen Guthery[14] · Marla C. Dubinsky[15] · Jason M. Shapiro[16] · Anthony R. Otley[17] · Mark Daly[5] · Lee A. Denson[4] · Subra Kugathasan[2] · Michael E. Zwick[1]

1 Department of Human Genetics, Emory University, Atlanta, GA, USA

2 Department of Pediatrics, Emory University, Atlanta, GA, USA

3 Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

4 Division of Pediatric Gastroenterology, Hepatology, and Nutrition, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

5 Broad Institute of MIT and Harvard, Cambridge, MA, USA

6 Division of Pediatric Gastroenterology, Hepatology and Nutrition, Department of Pediatrics, The Hospital for Sick Children, University of Toronto, Toronto, ON, Canada

7 Department of Pediatrics, University of Pennsylvania, Children's Hospital of Philadelphia, Philadelphia, PA, USA

8 Department of Pediatric Gastroenterology, Hepatology, and Nutrition, Medical College of Wisconsin, Milwaukee, WI, USA

9 Division of Digestive Diseases, Hepatology, and Nutrition, Connecticut Children's Medical Center, Hartford, CT, USA

10 Department of Pediatric Gastroenterology, Nationwide Children's Hospital, The Ohio State University College of Medicine, Columbus, OH, USA

11 Department of Pediatrics, The University of Chicago Comer Children's Hospital, Chicago, IL, USA

12 Department of Pediatrics, University of California at San Francisco, San Francisco, CA, USA

13 Department of Gastroenterology and Nutrition, Boston Children's Hospital, Boston, MA, USA

14 Department of Pediatrics, University of Utah, Salt Lake City, UT, USA

15 Department of Pediatrics, Mount Sinai Hospital, New York, NY, USA

16 Department of Pediatrics, Hasbro Children's Hospital, Providence, RI, USA

17 Department of Pediatrics, Dalhousie University, Halifax, NS, Canada