SOMATO PUBLICATIONS

Annals of Gastroenterology and Digestive Disorders

Research Article

Identification of Novel Genetics Loci Associated with Crohn's Disease Using Network Proximity

Sam Kara[1,2], Alaa Hanna[2], Conrad T. Gilliam[1] and George D. Wilson[2]

[1]University of Chicago, Departments of Human Genetics, 920 East 58th St., Chicago, Illinois 60637, USA

[2]Radiation Oncology Department, Beaumont Health, 3811 W Thirteen Mile Road, Royal Oak, Michigan 48073, USA

***Address for Correspondence:** George D. Wilson, William Beaumont Hospital, Department of Radiation Oncology, 3811 W Thirteen Mile Road, Royal Oak, Michigan 48073, USA, Tel: +1 2485510214; Fax: +1 2485512443; E-mail: george.wilson@beaumont.edu

Received: 07 May 2019; **Accepted:** 03 June 2019; **Published:** 04 June 2019

Citation of this article: Kara, S., Hanna, A., Gilliam, CT., Wilson, GD (2019) Identification of Novel Genetics Loci Associated with Crohn's Disease Using Network Proximity. Ann Gastroenterol Dig Dis, 2(2): 010-018.

ABSTRACT

Background: We investigated the hypothesis that known Crohn's disease (CD) loci can be used to inform the search for novel CD using "network proximity" within a whole genome network.

Methods: CD genes were selected from a meta-analysis solely on the basis of genetic evidence and independently of gene function. Ingenuity Pathway Analysis (IPA) was used to predict molecular interaction relationships among CD seed genes and to predict novel ones that are then tested in CD genome-wide association studies (GWAS) and in two novel CD-associated genes were studied in collection of 336 CD cases and 338 controls.

Results: We identified a group of 22 CD genes that are connected in a highly non-random manner within different whole genome networks. From these genes we were able to demonstrate that the subset of genes that show maximized connectivity among CD candidate genes are themselves enriched as targets for CD predisposing mutations. We identified two novel CD candidate genes, *TRIM15* and *GRB2* and studied them in a case-controlled study where we found the single nucleotide polymorphism (SNP) minor allele (G) frequency for *TRIM15* to be 0.34 in controls and 0.4 in cases (P =0.021) Likewise, the *GRB2* SNP generated a significant P-value (P =0.015).

Conclusions: Our findings suggest that genetic variants in *TRIM15* and *GBR2* are associated with CD, and that network candidate gene prediction data provide compelling evidence that network proximity within whole genome molecular interaction databases serves as a useful proxy for predicting multigene patterns of inheritance for CD.

Keywords: Crohn's Disease, Genome wide association studies, Network proximity

Introduction

Inflammatory bowel disease (IBD), which includes Crohn's Disease (CD) and ulcerative colitis (UC), is characterized by chronic intestinal inflammation. Although the exact origin of IBD remains unknown, evidence suggests that the interplay among immune response, environmental factors, and multiple genes are associated with susceptibility to IBD [1]. Epidemiologic data support genetic contribution to the pathogenesis of IBD, which include familial aggregation, twin studies, and racial and ethnic differences in disease prevalence. In recent years, most progress has been made in understanding the genetic contribution to disease pathogenesis. The currently known IBD risk loci showed an almost 75% overlap with genetic risk loci for other immune-mediated diseases [2]. Recently, we have shown that autoimmune diseases risk loci, including CD, cluster more proximal to each other among the total genome interaction network [3]. In addition, several new genes have been identified to be involved in the genetic susceptibility to CD [4]. The characterization of novel genes potentially will lead to the identification of therapeutic agents and clinical assessment of the phenotype and prognosis in patients with CD.

One of the major challenges in human genetics and marker development today is to understand the genetic architecture of disorders in humans. The techniques available for marker development are genomics and proteomics, and most recently bioinformatics. With the advent of powerful GWAS and large multi-site collaborative study designs it is possible to scan the entire genome for statistical associations between common variants and disease status. The central tenet of this study is that multigene contributors to heritable disorders will map more proximal to one another within whole genome molecular interaction databases. Consequently, network proximity which is a measure of the number of molecular interactions separating individual molecules within a molecular interaction network, can be used in combination with GWAS data to explore the genetic architecture of complex human disease.

CD is an ideal disease to test the potential of network proximity as it is characterized by high heritability [1,5,6]. To date there are many loci implicated in the genetic etiology of CD [7,8] and there is overlapping genetic etiology with other autoimmune-inflammatory conditions [9,10]. Previously published positional cloning or GWAS converge upon the prediction of more than 200 confirmed risk loci reported to contribute to the predisposition to CD [8,9,11-21], however few have been conclusively resolved to specific functional variants [22]. Recently, we have shown that a molecular interaction network approach (MINA) was able to identify novel genes which overlapped in seven autoimmune diseases, including CD [3]. Thus, CD provides an ideal test case to evaluate whether genome-wide MINA can be used to inform the search for multigenic determinants of IBD.

To test our hypothesis, we adopted a set of 22 candidate CD genes identified in a meta-analysis of three GWAS studies [8] and designated these CD "seed genes". We reasoned that this set of 22 genes share (or are enriched for) the common biological property that inherited alteration of DNA sequence in or around known coding-DNA segments that predisposes the bearer to CD. We used a commercial whole genome molecular interaction database developed in Ingenuity Pathway Analysis (IPA) to study the proximity of the seed genes. The goal was to test whether this set of 22 genes were located more proximal to one another in whole-genome molecular interaction databases than predicted by chance. Secondly, we investigated whether the minimal set of "connecting genes", required to link the maximal number of seed genes, would be enriched as targets for CD-predisposing mutations. We provide evidence to support of both these predictions.

Materials and Methods

Study design

We used "network proximity" to identify a small number of candidate genes that were then "re-evaluated" in the published GWAS studies [3]. By lowering the number of SNPs tested, we sought to detect candidate CD genes that were indistinguishable from background noise in the genome wide studies. We used this strategy to inspect a set of 17 candidate genes that could be evaluated in a novel case-controlled study. Thus, our study design is based on the identification and association analysis of a very small number of candidate genes (relative to a whole genome scan) where the statistical cost of multiple testing is greatly reduced.

Selection of Crohn's Disease candidate, or "seed" genes

Our goal was to select a threshold number of highly likely CD genes sufficient to seed the formation of a molecular interaction gene network which would be enriched for other CD candidate genes. The seed genes (Table 1) were selected based on a meta-analysis of three previously published GWAS studies [8] that combined 3,230 cases and 4,829 controls (and replication in 3,664 independent cases); all of which were European descent [19].

Molecular network analysis

We used the Ingenuity Pathway Analysis (IPA) software (Build 481437M, 2018) (Ingenuity Systems, Mountain View, CA, USA) and Knowledge Base to predict molecular interaction relationships among CD candidate genes and to predict novel CD genes. Seed genes were uploaded to the IPA. The IPA first searches for evidence of direct or indirect interaction between seed genes until the maximum number of seed genes are incorporated into the default 35-member network. The algorithm next seeks to connect the maximum number of remaining candidate genes allowing for one "connecting" gene between established and new members of the network.

Table 1: Top-ranking Molecular Interaction Networks Identified by the Ingenuity IPA Software.
The top-ranking networks are shown. For each network, the genes, the total number of genes, the score (see Materials and Methods), the seed genes, and the top-three ranked functions are listed.
*Crohn's disease seed genes are shown in Bold.

Network ID	Genes in Network	Number of Genes	Top-ranked Functions
1	ATG16L1, EMSY, CBX5, CCR6, CDKAL1, CLEC4C, CSF2, ESYT1, GRB2, HES5, HNF4A, IKBKG, IL12RB1, IL12RB2, IL3, IL12B, IL23A, IRGM, JAK2, LEP, LTB4R2, LY75, MST1, NKX2-3, NOD2,PEMT, PTGER4, PTPN2, PTPN22, STAT3, TMEM49, TNF, TNFSF15, TRIM15, ZNF365	35	Gastrointestinal Disease, Antigen Presentation, Cell-mediated Immune Response
2	LRRK2, PARK2	2	Post-Translational Modification, Genetic Disorder, Neurological Disease
3	ITLN1, LTF	2	Cell Morphology, Infection Mechanism, Molecular Transport
4	ESR1, ICOS, ICOSLG	3	Cell-mediated Immune Response, Humoral Immune Response, Lymphoid Tissue Structure and Development
5	EEF1A1, MIRN154, MIRN188, MIRN26A1, MIRN26A2, ORMDL3, XBP1	6	Cell Morphology, Cellular Assembly and Organization, Lipid Metabolism

Graphical representation of gene networks

We use the term 'network' to refer to a graphical representation of the molecular relationships between genes or gene products. Genes or gene products are represented as nodes/ shapes, and the biological relationship between two nodes is represented as an edge (line). In order to facilitate visualization of the seed and connecting genes we only show the molecular interactions (edges) connecting network members. IPA software was used to "link together the maximum number of seed genes with a minimal number of connecting genes within the constraints of the default 35-node network".

P-score and P-value calculations

IPA calculates the probability that the resulting network occurred by chance; i.e., that the final network is comprised of a random collection of genes. As shown in Table 1, a score of 2 indicates there is a 1/100 chance (P = 0.05; 99% confidence level) that the listed group of genes were incorporated randomly into the molecular interaction network. The *P*-value is calculated using Fisher's exact test. The *P*-score is defined as a *P*-score = $-\log_{10}$ (*P*-value). A right tailed Fisher's exact test (with a p = 0.05) was used to calculate the probability that each biological function assigned to two or more network genes occurred by chance.

Genome wide association studies in WTCCC-CD and NIDDK- IBD datasets

We present an approach for comparing genetic architecture of disease for which GWA data is available. Our approach relies on the raw summary statistics of GWAS and does not require obtaining individual level genotype data. The Wellcome Trust Case Control Consortium-Crohn's disease dataset (WTCCC-CD) dataset includes approximately 3,000 controls and 2000 unrelated individuals diagnosed with CD[10]. The National Institute of Digestive Disorders and Kidney disease – Inflammatory Bowel Disease (NIDDK- IBD) Genetics Consortium GWAS dataset includes 968 ileac Crohn's cases and 995 matched unrelated controls [11,23]. The data were sorted according to the chromosomal location of all genotyped SNPs. We cross referenced the SNPs and their chromosomal location (build 129) with the location of all known genes in the NCBI's gene dataset build 36. We identified a gene location as 50 Kilo-base pair (Kb) up- and down-stream of NCBI's start- and end- gene location [24], respectively. The number of independent SNPs for each gene was calculated [25]. The association results for all SNPs in the candidate genes were extracted and a gene *P*-values was obtained by multiplying the smallest *P*-value (for the given gene) with the total number of independent SNPs genotyped in each gene with the total number of tested genes (this corresponds to a Bonferroni correction).

The university of chicago TRIDOM patients and samples

Genomic DNA samples were obtained from the Translational Research Initiative in the Department of Medicine (TRIDOM) at the University of Chicago. A total of 376 CD patients (176 female, 198 male) were studied. Patients in the study were of European-American (n=336), black and African-American (n=19), Hispanic (n=4), Asian (n=3), and Indian-American (n=1) ancestries. All patients met the American Gastroenterological Association criteria for diagnosis of CD. The study was approved by the institutional review boards at all institutions, and informed consent was obtained from all subjects in the study. For this particular study, only European-American patients were considered. A total of 384 DNA samples of European-American ancestry and used previously as control individuals for the study of anxiety and related disorders were obtained from Columbia University and used as controls [26]. The observed genotype frequency in cases and controls did not deviate from Hardy-Weinberg proportions ($P > .01$).

SNP genotyping

The rs2517646 and rs4789182 SNPs associated with *TRIM15*and *GRB2*, respectively, were genotyped using the AgenaiPLEX™ assay and MassArray platform according to the manufacturer's protocols (Agena, San Diego, CA).

Statistical analysis

The genotype-phenotype genetic association *P*-values were extracted from the GWAS databases described above. We estimated the number of independent SNPs for each gene, using the method of Nyholt [26] which performs spectral decomposition of matrices of pairwise LD between SNPs. HapMap and 1000 genomes datasets were used to estimate LD. We applied a Bonferroni correction, based on the number of independent SNPs, to all uncorrected *P*-values and reported both the smallest uncorrected *P*-value for each gene along with the smallest corrected value. *P*-values less than 0.05, obtained by this method were, considered to be, significant. To evaluate evidence of genotype-phenotype association, we selected the smallest *P*-value in each gene and adjusted with a Bonferroni correction based on the number of independent SNPs. All reported *P*-values are adjusted *P*-values (unless indicated). The association between the *TRIM15* and *GRB2* SNPs and CD in the TRIDOM sample was investigated using a chi-square allelic test.

Results

Characterization of crohn's disease molecular networks

Of the twenty-two seed genes (Table 1), 21 were represented in the IPA molecular interaction database (MUC19 was omitted) and were used to generatea 35-member network. Figure 1 depicts the top-ranking network from this analysis. This 35-member network comprised of 17 "seed" genes and 18 "connecting" genes (Table 2). The Ingenuity database provides information regarding a given gene's cellular location and classification of genes by standard gene ontology classifications (Supplement table 1). A number of the seed genes, as well as the network connecting genes, are G-protein coupled receptors or growth factors with established roles in signaling, cellular development, cell growth and proliferation, and inflammatory response (Supplement table 1). The 35-gene network contained 28 genes directly connected and 8 genes indirectly connected (5 seed genes; NKX2-3, PTGER4, ZNF365, TNFSF15, and CCR6 and two connecting genes; LY75 and CLEC4C). Supplement table 2 list all the 35 member genes, their interactions, and their interaction type. We next asked whether the incorporation of 17 out of 21 "seed" genes into a single 35-member network is likely to have occurred by chance. The Ingenuity software generates a statistical score for each developed network expressed as the likelihood that the generated network occurred by chance. Our top network showed a score of 48 indicating that the software would produce this particular network by chance only once in every 10^{48} simulations.

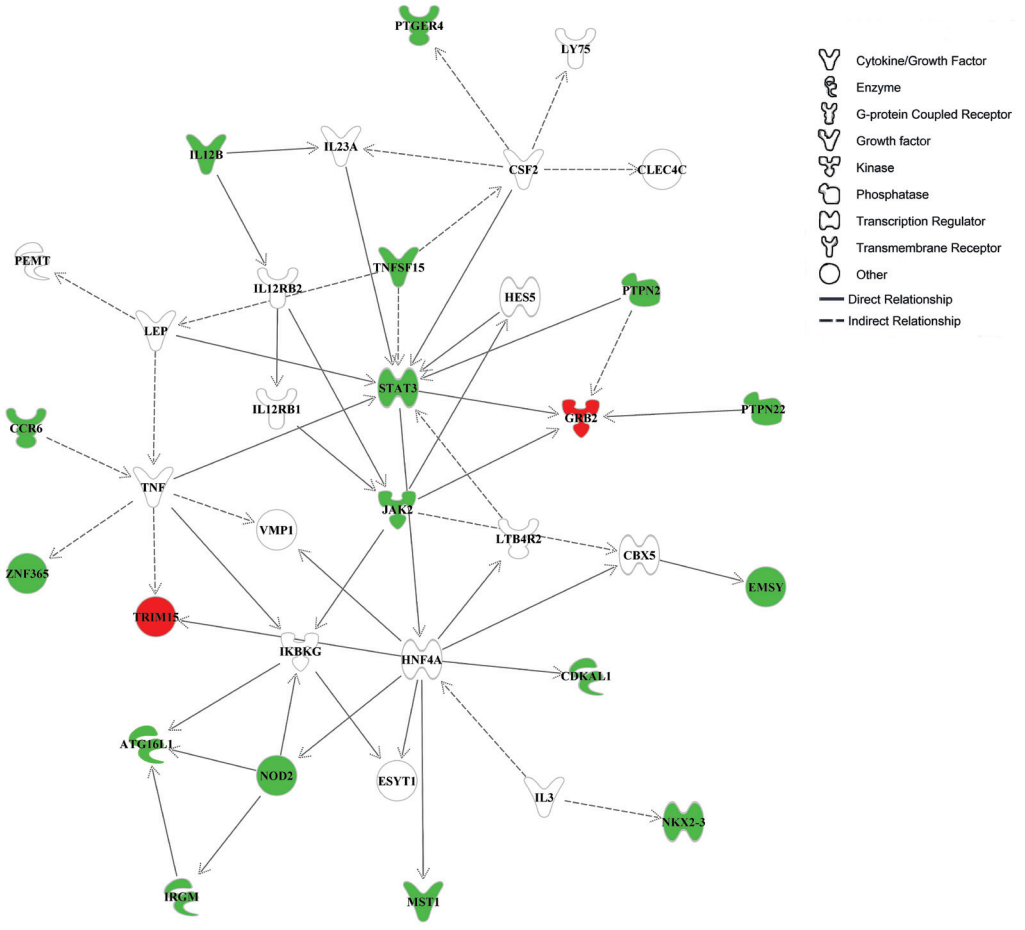

Figure 1: Schematic diagram of a Crohn's disease gene-enriched molecular interaction network.
Putative Crohn's disease genes, or seed genes, are shown in grey; connecting genes are depicted as open figures. Solid lines represent evidence for "direct" interaction between the connected genes (i.e, bind, cleave, etc.); dotted lines signify "indirect" interactions (e.g. activate, inhibit, etc). Interactions between network genes, and non-network genes are not shown. The IPA software includes information (for some genes) about the subcellular localization of each gene product. Gene functions are inferred either from experimental data or from gene ontology analysis.

Association analysis of network connecting genes in publicly available GWAS datasets

18 of the final 21 CD "seed" genes showed significant association to CD in the WTCCC database (corrected $P = 0..05$) including seven genes with *P*-values less than 0.00001: (*NOD2*; *IL23A*; *ATG16L1*; *MST1*; *IRGM*; *NKX2-3*; and *PTPN2*). Among these seven genes, *NOD2*, *IL23A*, and *ATG16L1* generate similar statistical evidence of association in the NIDDK-IBD dataset (data not shown).Table 2 lists the 18 connecting network genes along with their chromosomal locations and association results for the previously published and publicly available WTCCC-CD datasets. We tallied all SNPs within and immediately surrounding each of the candidate genes, estimated the number of independent SNPs from the HapMap LD structure, and for each gene we adjusted the genotype-phenotype correlation scores for multiple testing using a conservative Bonferroni correction. Table 2 lists the SNP ID corresponding to the smallest *P*-value detected per gene in each study, the effective number of independent SNPs genotyped in each gene, and the corrected *P*-values for the highest-ranking SNP. Under the null hypothesis of no association we expect on average less than one gene to demonstrate an adjusted P = 0.05.

As shown in Table 2, all of the genes (except for *HES5* and *CLEC4C*) showed association to CD before multiple testing adjustments. Several of these genes have been implicated as CD genes in previous studies (interleukin 12 receptor, beta 2 (*IL12RB2*), *TNF*, *IL12RB1*, and *HNFA4*) but were not selected as seed genes for the current study. For example, SNP rs12119179 from the *IL12RB2* is strongly associated with CD after multiple testing adjustment ($P = 0.7.94E10^{-09}$). Association between *IL12RB2* and CD was reported in the initial publication of the WTCCC-CD data [10] and more recently in a report using pathway analysis to identify susceptibility genes and gene-gene interactions [27-30]. After multiple testing adjustments, analysis of the WTCCC-CD data implicates *IL12RB2* and four other genes, *CSF2*, *IL3*, *TRIM15*, and *GRB2*, that we propose as new CD candidate genes. We report novel evidence for CD-association with

Table 2: Candidate genes in the CD-specific network and their association in the WTCCC-CD dataset.
"a" No data were available. The candidate genes and their chromosomal positions (Chr. Position) are shown. For each gene, the SNP (SNP ID) with the smallest association *P*-value and the total numbers of independent SNPs investigated for each gene (SNP count) are shown in the WTCCC-CD dataset. The Bonferroni's adjusted *P*-values for the best SNP within each gene (adjusted *P*-value; the total number of independent SNPs tested in all selected genes), are also presented. Genotype counts, frequency, and association are available from www.wtccc.org.uk

Gene	Chr. Position	SNP ID	Smallest *P*-value	Independent SNP count	Adjusted *P*-value
HES5	1p36.32	rs2495368	0.066	5	1
IL12RB2	1p31.3-p31.2	rs12119179	5.40E-11	11.21	7.94E-09
LY75	2q24	rs2729706	0.036	17.16	1
CSF2	5q31.1	rs27437	0.00032	7.05	0.047
IL3	5q31.1	rs27437	0.00032	7.06	0.047
TNF	6p21.3	rs2736172	0.027	8	1
TRIM15	6p21.3	rs2517646	0.000034	10	0.004
LEP	7q31.3	rs10954177	0.0047	17.05	0.69
CBX5	12q13.13	rs10506328	0.0248	3.31	1
CLEC4C	12p13.2-p12.3	rs10845821	0.1	6	1
ESYT1	12q13	rs2292239	0.044	6	1
LTB4R2	14q11.2-q12	rs1950501	0.00077	7	0.11
GRB2	17q24-q25	rs16967789	0.00024	6	0.03
PEMT	17p11.2	rs8074272	0.0146	8.15	1
TMEM49	17q23.1	rs1296279	0.00491	9.01	0.72
IL12RB1	19p13.1	rs419540	0.0051	6	0.74
HNF4A	20q12-q13.1	rs3181206	0.0327	13	1
IKBKG'a'	Xq28				

the tripartite motif-containing 15 (*TRIM15*) gene (rs2517646; P = 0.004); a single SNP (rs27437; P = 0.047) implicating two adjacent genes on chromosome 5, the colony stimulating factor 2 (granulocyte-macrophage) (*CSF2*) gene and the interleukin 3 (*IL3*) gene; and the Growth factor receptor-bound protein 2 (*GRB2*) on chromosome 17 (rs16967789; P = 0.03).

Allelic association of the connecting genes in GWAS studies

We next asked whether the 18 connecting genes detect allelic association to CD in the publicly available NIDDK-IBD GWAS dataset. Several candidate genes showed significant association values before adjustment (data not shown), however only *IL12RB2* remains statistically significant (rs10889677; P = $01.6x10^{-9}$). Not finding evidence for replication in the NIDDK-IBD dataset, we attempted to replicate our findings in a novel collection of 336 CD cases and 338 controls. Given the relatively small size of the independent sample we sought to increase our chance of detecting association to the disease phenotype by decreasing the number of independent tests performed. Comparison of the 30 independent SNPs spanning the four candidate genes (*IL3*, *CSF2*, *TRIM15* and *GRB2*) revealed 2 SNPs with the strongest support for association to disease phenotype in both the WTCCC and NIDDK datasets; the rs2517546 allele of *TRIM15* and the rs4789182 SNP in *GRB2* (as shown in Table 2, *GRB2* SNP rs16967789 was the single most significant finding, however it was not genotyped in the NIDDK-IBD dataset. SNP rs4789182 was genotyped in both datasets and maps in strong disequilibrium with rs16967789). Under the less stringent multiple test correction afforded by the analysis of only 30 independent tests the *TRIM15* SNP generated adjusted *P*-values of 0.001 and 0.009 in the WTCCC-CD and NIDDK-IBD datasets, respectively. The *GRB2* SNP showed an adjusted *P*-value of 0.027 in the WTCCC-CD dataset and 0.03 in the NIDDK-IBD dataset after adjustment for multiple testing under these conditions.

Analysis of TRIM15 and GRB2 in a novel crohn's disease case-control sample

We evaluated the *TRIM15* SNP (rs2517546) and a *GRB2* SNP (rs4789182) in TRIDOM; a novel case-control sample of 336 European-American CD cases and 384 controls. For *TRIM15*, the SNP minor allele (G) frequency was found to be 0.34 in controls and 0.4 in cases (Table 3). This finding is consistent with association to CD (P = 0.021) in the TRIDOM cohort and thus further validates the proposal that *TRIM15*, or possibly a nearby gene, is a CD-associated gene. Likewise, the *GRB2* SNP generates a significant *P*-value (P = 0.015) and adds to the evidence that *GRB2* is implicated in CD etiology (Table 3).

Discussion

One of the persistent challenges in the genetic study of common heritable disorders like CD is that individual genotype-phenotype correlations tend to be weak and consequently the distinction between real and false positive correlations tend to be unreliable. We recently reported the application of molecular interaction network proximity in predicting genetic inheritance of autoimmune disease [3], like CD. We selected seed genes solely on the basis of GWAS data in order to minimize inherent biases that can arise when drawing inferences from molecular interaction data. With this caveat in mind,

Table 3: Association analysis in a novel sample of 336 CD cases and 384 control subjects.
The individual SNPs showing strongest evidence for allelic association with Crohn's disease in the WTCCC-CD and NIDDK-IBD datasets - *TRIM15* (rs2517546) and *GRB2* (rs4789182) - were genotyped in a novel sample of 336 CD cases and 384 control subjects.

SNP id	Allele	AA	AB	BB	Count	*P*-value
rs2517646	Controls	39	154	149	342	**0.021**
	Cases	54	159	121	334	
rs4789182	Controls	140	187	43	370	**0.015**
	Cases	166	133	37	336	

we selected the study of CD based on the relatively large number of reported candidate genes and the relatively large proportion of these genes that have been identified in two or more independent GWAS studies [8]. All of the seed genes selected for this study have been independently reported in two or more GWAS studies and have been mapped to 30 previously reported CD risk loci [8]. The most straightforward question we asked was whether the list of candidates "seed" genes are mapped more proximal to one another than predicted by chance in molecular interaction databases. Network proximity can presumably arise from shared biological properties or biological roles, but also from other, potentially confounding factors including: the order in which nodes (i.e, proteins) evolved [31-33], inherent 'scale-free' properties of biological networks (i.e., failure to correct for the contribution of hub nodes) [34,35], or other artifacts that can arise in the generation of the molecular networks. Clearly, the generation of false positive evidence for network proximity is a potential confound.

We used IPA and whole genome molecular interaction database to evaluate network proximity among the CD candidate genes. The IPA analytics provided a statistic (network score) to estimate the likelihood that our observation (identification of a single 35-member network that includes 17 of the original 22 seed genes) could have occurred by chance (network score of 48= P = 010^{-48}). While this is an overwhelming significance value, the statistical algorithm was not developed specifically for our purposes and thus we could not be certain of its relevance. We have described the reproducibility of the IPA recently [3].

A related but more challenging question was whether the minimum number of genes (connecting genes) required to connect the maximum number of seed genes is enriched for genetic variation that predisposes individuals to CD. The latter is challenging because there is no reliable way to distinguish false positive genetic association from real association when the genotype-phenotype correlation is relatively weak, as have been the vast majority of correlations for common genetic disorders to date.

Using the IPA platform, we provide evidence that5 of the 18 connecting genes are implicated in CD. Among these connecting genes *CSF* and *IL3* reside in a frequently replicated region associated with CD [7,8] and *IL12RB2* resides near the well-established CD gene *IL23A*. While we make no claims that these are novel findings, they do serve to validate the network proximity methodology. For the two novel candidate genes, *TRIM 15* and *GRB2*, we provide additional evidence of their association to CD in an independent case: control association study. Taken together, we argue the network simulation data and candidate gene prediction data provide compelling evidence that network proximity within whole genome molecular interaction databases serves as a useful proxy for predicting multigene patterns of inheritance for CD, and my inference, other common heritable disorders.

Based on these findings it seems likely that judicious application of molecular interaction data to GWAS data will prove useful in the identification of novel complex trait loci and thereby increase the likelihood that biologically relevant interactions or pathways will be identified and subsequently targeted for clinical interventions. Whether the use of molecular interaction networks will help identify epistatic interactions that predict significant population-based disease susceptibility remains unknown. In a recent study of behavioral variation in *Drosophila*, where genetic background and environment are greatly simplified relative to human studies, and where random genetic variation is replaced by insertional mutagenesis, investigators used molecular interaction networks to provide compelling evidence that changes in epistatic gene-gene interaction – as opposed to differences in single gene variants - account for a majority of the genetic attributable risk for particular behavioral phenotypes in flies [36]. On the other hand, the precise effects of specific genetic variants on bristle number in inbred flies have been mapped to several genes, and yet those same variants within a natural population have little bearing on bristle counts [37].

Apart from the broader implications discussed above, whole-genome molecular interaction data is destined to capture information that would not likely be available otherwise. The single most statistically significant finding from these studies is the identification of *TRIM15* as a novel CD candidate gene in the WTCCC-CD and NIDDK-IBD studies, and in a novel case: control sample. *TRIM15* is a good example of a CD candidate gene whose discovery would have been unlikely without the use of large-scale molecular interaction networks. We were unable to identify biological interactions between *TRIM15* and other seed or connecting genes using manual search strategies, and whereas most of the other connecting genes identified by the IPA software and interaction database could be replicated using the GeneWays system [32], the GeneGoplatform, or other publicly available systems like *STRING* [38] (data not shown), only the IPA system links *TRIM15* to the CD network. Follow up of the IPA predictions revealed that the direct interaction depicted in Figure 1 between *TRIM15* and *HNF4A* was documented in the Supplement Section of a manuscript published in Science in 2004 [39]. The predicted indirect interaction traces back to a 2003 manuscript [40] reporting that the "TNF protein increases expression of human ZNFB7 [*TRIM15*] mRNA".

TRIM15 maps to the major histocompatibility locus (MHC) region on chromosome 6p21.3 which has been implicated in autoimmune diseases, and gene Ontology studies suggested its role in innate immune response, and CD by both genetic linkage [16] and association [41,42] studies. The MHC region is characterized by long-ranging LD spanning several immune-active genes, so that pinpointing a locus for CD has presented a significant challenge [7,43]. The *TRIM15* gene maps 1.4Mb distal to *TNF* which was also

identified by our network studies and which has long been implicated in CD [42]. Using the HapMap/ 1000 genomes datasets, we found no evidence for LD between *TRIM15* and TNF. The human *TRIM15* gene is a non-class I gene that plays an important defensive and regulatory role as part of the immune response system in disease and infection [44]. It has been noted recently that a growing number of genes harbor predisposition alleles for multiple autoimmune disorders [45-48]. We note in this regard that independent studies implicate *TRIM15* in the predisposition to autoimmune disease such as systemic lupus erythematous, Insulin-dependent diabetes mellitus, rheumatoid arthritis, Multiple Sclerosis, and colon cancer [29,44,49-53]. Furthermore, during fetal development, *TRIM15* was reported to be expressed, almost exclusively, in intestine starting at 10 weeks and peaking at 20 weeks, with minor expression in the stomach at 10 weeks and declining to background level at 20 weeks [54]. *TRIM15* RNA-sequencing of 27 different tissues showed biased expression in duodenum, small intestine, colon, and minute expression level in six other tissues [55].

We also report evidence for genetic association between the Growth factor receptor-bound protein 2 (*GRB2*) and CD in our re-analysis of the WTCCC-CD dataset. *GRB2* is an adaptor protein that provides a critical link between cell surface growth factor receptors and the Ras signaling pathway [56]. *GRB2* is located in the cytoplasm along with a number of other CD seed genes (*IRGM*, *JAK2*, *PTPN22*, *PTP2*, *ATG16L1*, and *NOD2*), all of which are implicated by gene ontology analysis in signaling and interactions ($P = 04.8x10^{-10}$), inflammatory response ($P = 04.8x10^{-10}$) and immunological disease ($P = 04.64x10^{-8}$). *GRB2*, as well as *CSF2* and *IL3* map alongside other seed genes and established CD candidate genes (*IL12B*, *IL23A*, *JAK2*, *STAT3*, *IL12RB1*, and *IL12RB2*) in the JAK-STAT signaling pathway. There is evidence that the type III inflammatory responses typically seen in rheumatoid arthritis, and other autoimmune diseases such as irritable bowel disease, chronic heart failure, and diseases of the eye, accelerate when the JAK/STAT pathway itself becomes dysregulated [57]. We note that a recent meta-analysis mapped over 35 genes of interest to 71 confirmed CD loci [7]. This analysis added to the number of CD-implicated genes but did not identify *TRIM15* and *GRB2*.

In summary, using our novel network approach, we provide evidence of genetic factors influencing the risk of CD, with corroboration of our findings (e.g., the association of *TRIM15*and *GRB2*in the NIDDK-IBD, WTCCC-CD, and TRIDOM) needed in future studies.

Acknowledgements

This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113.

"The NIDDK IBD Crohn's Disease Genome-Wide Association Study was conducted by the NIDDK IBD Crohn's Disease Genome-Wide Association Study Investigators and supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). This manuscript was not prepared in collaboration with Investigators of NIDDK IBD Crohn's Disease Genome-Wide Association Study and does not necessarily reflect the opinions or views of the NIDDK IBD Crohn's Disease Genome-Wide Association Study or the NIDDK."

Funding

This work was supported by generous donations from the following: David and Janice Katz, Judy McCormack, the Ryan Licht Sang Bipolar Foundation, and Debbie and Larry Hilibrand.

References

1. Podolsky, DK. (2002) Inflammatory bowel disease. N Engl J Med, 347: 417-429.

2. Spekhorst, LM., Visschedijk, MC., Weersma, RK., Festen, EA. (2015) Down the line from genome-wide association studies in inflammatory bowel disease: the resulting clinical benefits and the outlook for the future. Expert Rev Clin Immunol, 11(1): 33-44.

3. Kara, S., Pirela-Morillo, GA., Gilliam, CT., Wilson, GD. (2019) Identification of novel susceptibility genes associated with seven autoimmune disorders using whole genome molecular interaction networks. J Autoimmun, 97: 48-58.

4. Verstockt, B., Smith, KG., Lee, JC. (2018) Genome-wide association studies in Crohn's disease: Past, present and future. Clin Transl Immunology, 7(1): e1001.

5. Stracke, S., Haseneyer, G., Veyrieras, JB., Geiger, HH., Sauer, S., Graner, A., et al. (2009) Association mapping reveals gene action and interactions in the determination of flowering time in barley. TAG Theoretical and applied genetics Theoretische und angewandte Genetik, 118(2): 259-273.

6. Weersma, RK., Stokkers, PC., van Bodegraven, AA., van Hogezand, RA., Verspaget, HW., de Jong, DJ., et al. (2009) Molecular prediction of disease risk and severity in a large Dutch Crohn's disease cohort. Gut, 58(3): 388-395.

7. Franke, A., McGovern, DP., Barrett, JC., Wang, K., Radford-Smith, GL., Ahmad, T., et al. (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat genet 42(12): 1118-1125.

8. Barrett, JC., Hansoul, S., Nicolae, DL., Cho, JH., Duerr, RH., Rioux, JD., et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat genet, 40(8): 955-962.

9. Cargill, M., Schrodi, SJ., Chang, M., Garcia, VE., Brandon, R., Callis, KP., et al. (2007) A large-scale genetic association study confirms IL12B and leads to the identification of IL23R as psoriasis-risk genes. Am J Hum Genet, 80(2): 273-290.

10. Wellcome Trust Case Control C. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 447(7145): 661-678.

11. Rioux, JD., Xavier, RJ., Taylor, KD., Silverberg, MS., Goyette, P., Huett, A., et al. (2007) Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat genet, 39(5): 596-604.

12. Raelson, JV., Little, RD., Ruether, A., Fournier, H., Paquin, B., Van Eerdewegh, P., et al. (2007) Genome-wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci. Proc Natl Acad Sci U S A, 104(37): 14747-14752.

13. Parkes, M., Barrett, JC., Prescott, NJ., Tremelling, M., Anderson, CA., Fisher, SA., et al. (2007) Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat genet, 39(7): 830-832.

14. Newman, WG., Zhang, Q., Liu, X., Amos, CI., Siminovitch, KA. (2009) Genetic variants in IL-23R and ATG16L1 independently predispose to

increased susceptibility to Crohn's disease in a Canadian population. Journal of clinical gastroenterology, 43(5): 444-447.

15. Mathew, CG. (2008) New links to the pathogenesis of Crohn disease provided by genome-wide association scans. Nat Rev Genet, 9(1): 9-14.

16. Hampe, J., Shaw, SH., Saiz, R., Leysens, N., Lantermann, A., Mascheretti, S., et al. (1999) Linkage of inflammatory bowel disease to human chromosome 6p. Am J human Genet, 656(): 1647-1655.

17. Hampe, J., Franke, A., Rosenstiel, P., Till, A., Teuber, M., Huse, K., et al. (2007) A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nat genet 39(2): 207-211.

18. Franke, A., Hampe, J., Rosenstiel, P., Becker, C., Wagner, F., Hasler, R., et al. (2007) Systematic association mapping identifies NELL1 as a novel IBD disease gene. PloS one, 2(8): e691.

19. Duerr, RH., Taylor, KD., Brant, SR., Rioux, JD., Silverberg, MS., Daly, MJ., et al. (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science, 314(5804): 1461-1463.

20. Liu, JZ., van Sommeren, S., Huang, H., Ng, SC., Alberts, R., Takahashi, A., et al. (2015) Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat genet, 47(9): 979-986.

21. de Lange, KM., Moutsianas, L., Lee, JC., Lamb, CA., Luo, Y., Kennedy, NA., et al. (2017) Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat genet, 49(2): 256-261.

22. Huang, H., Fang, M., Jostins, L., Umicevic Mirkov, M., Boucher, G., Anderson, CA., et al. (2017) Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature, 547(7662): 173-178.

23. Libioulle, C., Louis, E., Hansoul, S., Sandor, C., Farnir, F., Franchimont, D., et al. (2007) Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PLoS genet, 3(4): e58.

24. Abecasis, GR., Cherny, SS., Cookson, WO., Cardon, LR. (2002) Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat genet, 30(1): 97-101.

25. Pailhez, G., Bulbena, A., Fullana, MA., Castano, J. (2009) Anxiety disorders and joint hypermobility syndrome: the role of collagen tissue. Gen Hosp Psychiatry, 31(3): 299.

26. Nyholt, DR. (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Human Genet,74(4): 765-769.

27. Zhang, W., Hui, KY., Gusev, A, Warner, N., Ng,SM,, Ferguson, J., et al. (2013) Extended haplotype association study in Crohn's disease identifies a novel, Ashkenazi Jewish-specific missense mutation in the NF-kappaB pathway gene, HEATR3. Genes Immun, 14(5): 310-316.

28. Lee, YH., Song, GG. (2012) Pathway analysis of a genome-wide association study of ileal Crohn's disease. DNA Cell Biol, 31(10): 1549-1554.

29. Lee, OH., Lee, J., Lee, KH., Woo, YM., Kang, JH., Yoon, HG., et al. (2015) Role of the focal adhesion protein TRIM15 in colon cancer development. Biochim biophys acta, 1853(2): 409-421.

30. Ballard, D., Abraham, C., Cho, J., Zhao, H. (2010) Pathway analysis comparison using Crohn's disease genome wide association studies. BMC medical genomics, 3: 25.

31. Rzhetsky, A., Wajngurt, D., Park, N., Zheng, T. (2007) Probing genetic overlap among complex human phenotypes. Proc Natl Acad Sci U S A, 104(28): 11694-11699.

32. Rzhetsky, A., Iossifov, I., Koike, T., Krauthammer, M., Kra, P., Morris, M., et al. (2004) GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. J Biomed Inform, 37(1): 43-53.

33. Deepak, P., Park, SH., Ehman, EC., Hansel, SL., Fidler, JL., Bruining, DH., et al. (2017) Crohn's disease diagnosis, treatment approach, and management paradigm: what the radiologist needs to know. Abdom Radiol, 42(): 1068-1086.

34. Barabasi, AL., Oltvai, ZN. (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet, 5(2): 101-113.

35. Barabasi, AL. (2009) Scale-free networks: a decade and beyond. Science, 325(5939): 412-413.

36. Yamamoto, A., Zwarts, L., Callaerts, P., Norga, K., Mackay, TF., Anholt, RR. (2008) Neurogenetic networks for startle-induced locomotion in Drosophila melanogaster. Proc Natl Acad Sci U S A, 105(34): 12393-12398.

37. Macdonald, SJ., Pastinen, T., Long, AD. (2005) The effect of polymorphisms in the enhancer of split gene complex on bristle number variation in a large wild-caught cohort of Drosophila melanogaster. Genetics, 171(4): 1741-1756.

38. Jensen, LJ., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., et al. (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res, 7: D412-416.

39. Odom, DT., Zizlsperger, N., Gordon, DB., Bell, GW., Rinaldi, NJ., Murray, HL., et al. (2004) Control of pancreas and liver gene expression by HNF transcription factors. Science, 303(5662): 1378-1381.

40. Pedron, T., Thibault, C., Sansonetti, PJ. (2003) The invasive phenotype of Shigella flexneri directs a distinct gene expression pattern in the human intestinal epithelial cell line Caco-2. J Biol Chem, 278(36): 33878-33886.

41. Tariq, A., Huengsberg, M., Cook, A., Ross, JD. (2002) Audit of official STD returns from genitourinary medicine. International journal of STD & AIDS, 13(10): 720-721.

42. Fidder, HH., Heijmans, R., Chowers, Y., Bar-Meir, S., Avidan, B., Pena, AS., et al. (2006) TNF-857 polymorphism in Israeli Jewish patients with inflammatory bowel disease. Int J Immunogenet, 33(2): 81-85.

43. Fisher, SA., Tremelling, M., Anderson, CA., Gwilliam, R., Bumpstead, S., Prescott, NJ., et al. (2008) Genetic determinants of ulcerative colitis include the ECM1 locus and five loci implicated in Crohn's disease. Nature Genet, 40(6): 710-712.

44. Ando, A., Shigenari, A., Kulski, JK., Renard, C., Chardon, P., Shiina, T., et al. (2005) Genomic sequence analysis of the 238-kb swine segment with a cluster of TRIM and olfactory receptor genes located, but with no class I genes, at the distal end of the SLA class I region. Immunogenetics, 57(11): 864-873.

45. Lettre, G., Rioux, JD. (2008) Autoimmune diseases: insights from genome-wide association studies. Hum Mol Genet, 17(R2): R116-121.

46. Bentham, J., Morris, DL., Graham, DSC., Pinder, CL., Tombleson, P., Behrens, TW., et al. (2015) Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat Genet, 47(12): 1457-1464.

47. International Consortium for Systemic Lupus Erythematosus, G., Harley, JB., Alarcon-Riquelme, ME., Criswell, LA., Jacob, CO., Kimberly, RP., et al. (2008) Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. Nat Genet, 40(2): 204-210.

48. Vyse, TJ., Richardson, AM., Walsh, E., Farwell, L., Daly, MJ., Terhorst, C.,

et al. (2005) Mapping autoimmune disease genes in humans: lessons from IBD and SLE. Novartis Found Symp, 267: 94-107.

49. Mukherjee, S., Guha, S., Ikeda, M., Iwata, N., Malhotra, AK., Pe'er, I., et al. (2014) Excess of homozygosity in the major histocompatibility complex in schizophrenia. Human Mol Genet, 23(22): 6088-6095.

50. Uchil, PD., Hinz, A., Siegel, S., Coenen-Stass, A., Pertel, T., Luban, J., et al. (2013) TRIM protein-mediated regulation of inflammatory and innate immune signaling and its association with antiretroviral activity. J Virol, 87(): 257-272.

51. Plenge, RM., Seielstad, M., Padyukov, L., Lee, AT., Remmers, EF., Ding, B., et al. (2007) TRAF1-C5 as a risk locus for rheumatoid arthritis--a genomewide study. N Engl J Med, 357(12): 1199-1209.

52. Matesanz, F., Gonzalez-Perez, A., Lucas, M., Sanna, S., Gayan, J., Urcelay, E., et al. (2012) Genome-wide association study of multiple sclerosis confirms a novel locus at 5p13.1. PloS one, 7(5): e36140.

53. Lanata, CM., Nititham, J., Taylor, KE., Chung, SA., Torgerson, DG., Seldin, MF., et al. (2018) Genetic contributions to lupus nephritis in a multi-ethnic cohort of systemic lupus erythematous patients. PloS one, 13: e0199003.

54. Szabo, L., Morey, R., Palpant, NJ., Wang, PL., Afari, N., Jiang, C., et al. (2016) Erratum to: Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. Genome Biol, 17(1): 263.

55. Fagerberg, L., Hallstrom, BM., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., et al. (2014) Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol Cell proteomics,13(2): 397-406.

56. Dharmawardana, PG., Peruzzi, B., Giubellino, A., Burke, TR., Jr., Bottaro, DP. (2006) Molecular targeting of growth factor receptor-bound 2 (Grb2) as an anti-cancer strategy. Anticancer drugs, 17(1): 13-20.

57. Malemud, CJ. (2009) The discovery of novel experimental therapies for inflammatory arthritis. Mediators Inflamm, 2009: 698769.