Manuscript Doi: 10.1093/ecco-jcc/jjac059

# Network analysis of inflammatory bowel disease research: towards the interactome

M. Emilia Fernandez[1], F. Nicolas Nazar[2†], Luciana B. Moine[1], Cristian E. Jaime[1], Jackelyn M. Kembro[2,3,*], Silvia G. Correa[1,4,*]

[1]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Centro de Investigaciones en Bioquímica Clínica e Inmunología (CIBICI), Córdoba, Argentina.

[2]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Instituto de Investigaciones Biológicas y Tecnológicas (IIByT), Córdoba, Argentina; Universidad Nacional de Córdoba, Facultad de Ciencias Exactas, Físicas y Naturales, Instituto de Ciencia y Tecnología de los Alimentos (ICTA), Córdoba, Argentina.

[3]Universidad Nacional de Córdoba, Facultad de Ciencias Exactas, Físicas y Naturales, Cátedra de Química Biológica, Córdoba, Argentina.

[4]Universidad Nacional de Córdoba, Facultad de Ciencias Químicas, Departamento de Bioquímica Clínica, Inmunología, Córdoba, Argentina.

Email addresses: mariaemilia.fernandez@mi.unc.edu.ar; franco.nicolas.nazar@unc.edu.ar; lu_moine@hotmail.com; cristian.jaime@mi.unc.edu.ar; jkembro@unc.edu.ar; silviagracielacorrea@unc.edu.ar.





† Present address: Department of Animal Production, NEIKER-Basque Institute for Agricultural Research and Development, Vitoria-Gasteiz, Spain

Manuscript Doi: 10.1093/ecco-jcc/jjac059

*Corresponding

**Jackelyn M. Kembro**, PhD, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Instituto de Investigaciones Biológicas y Tecnológicas (IIByT), Universidad Nacional de Córdoba, Facultad de Ciencias Exactas, Físicas y Naturales, Instituto de Ciencia y Tecnología de los Alimentos (ICTA) and Cátedra de Química Biológica, Av. Vélez Sarsfield 1611, X5000HUA, Córdoba, Argentina, Tel.: +54 351 535-3800, email: *jkembro@unc.edu.ar.*

**Silvia G. Correa**, PhD, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Centro de Investigaciones en Bioquímica Clínica e Inmunología (CIBICI), Universidad Nacional de Córdoba, Facultad de Ciencias Químicas, Departamento de Bioquímica Clínica, Inmunología, Av. Medina Allende y Haya de la Torre, Ciudad Universitaria X5000HUA, Córdoba, Argentina, Tel.: +54 351 535-3850, email: *silviagracielacorrea@unc.edu.ar.*

**Abbreviations**

ATG16L1: autophagy-related protein 16-1.

CARD: card family.

CASP, CASP3: caspase family, caspase 3.

CCL20: chemokine ligand 20.

CCR6: chemokine receptor 6.

CD: Crohn's disease.

Manuscript Doi: 10.1093/ecco-jcc/jjac059

FAS: fas cell surface death receptor.

GWAS: genome-wide association studies.

IBD: inflammatory bowel diseases.

IFN, IFNγ, IFNα: Interferon, Interferon gamma, Interferon alpha.

Ig, IgA, IgG: immunoglobulins, immunoglobulin A, immunoglobulin G.

IL: interleukins.

IRGM: immunity-related GTPase M.

NF-kB: nuclear factor kB.

NLRC4: NLR family CARD domain-containing protein 4.

NOD2: nucleotide-binding oligomerization domain containing protein 2.

NOS: nitric oxide synthase.

PCR: polymerase chain reaction.

PTPN2: tyrosine-protein phosphatase non-receptor type 2.

Q1, Q2, Q3, Q4: first, second, third and fourth quartile of the distribution.

QTL: quantitative trait locus.

ROS: reactive oxygen species.

SNP: single-nucleotide polymorphism.

Manuscript Doi: 10.1093/ecco-jcc/jjac059

STAT, STAT1, STAT3, STAT4: signal transducer and activator of transcription, signal transducer and activator of transcription 1, 3 and 4.

TGF, TGF-B: transforming growth factor, transforming growth factor beta.

TLR4: toll-like receptor 4.

TNF, TNFSF15: tumor necrosis factor, tumor necrosis factor ligand superfamily member 15.

UC: ulcerative colitis.

Extended non-standardized abbreviations can be consulted in Supporting Tables S4-6.

Manuscript Doi: 10.1093/ecco-jcc/jjac059


## Abstract

**Background and Aims:** Modern views accept that Inflammatory Bowel Diseases (IBD) emerge from complex interactions among the multiple components of a biological network known as "IBD interactome". These diverse components belong to different functional levels including cells, molecules, genes and biological processes. This diversity can make it difficult to integrate available empirical information in human patients into a collective view of etiopathogenesis, a necessary step to understand the interactome. Herein, we quantitatively analyze how representativeness of components involved in human IBD and their relations have changed over time.

**Methods:** A bibliographic search in PubMed retrieved 25971 abstracts of experimental studies on IBD in humans, published between 1990 and 2020. Abstracts were scanned automatically for 1218 IBD interactome components proposed in recent reviews. The resulting databases are freely available and were visualized as networks indicating the frequency in which different components are referenced together within each abstract.

**Results:** As expected, over time there was an increase in components added to the IBD network and heightened connectivity within and across functional levels. However, certain components were consistently studied together forming preserved motifs in the networks. These overrepresented and highly linked components reflect main "hypotheses" in IBD research in humans. Interestingly, 82% of the components cited in reviews were absent or showed low frequency, suggesting that many aspects of the proposed IBD interactome still possess weak experimental support in humans.

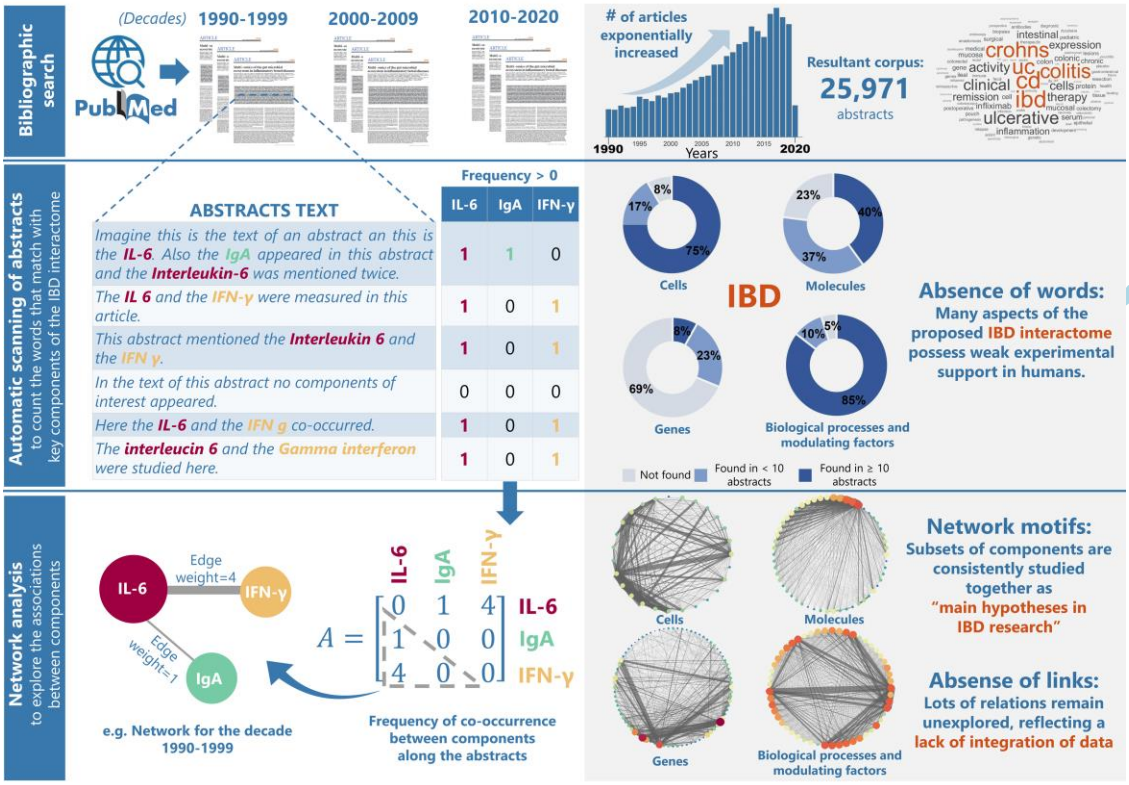
Manuscript Doi: 10.1093/ecco-jcc/jjac059

**Conclusions:** A reductionist and fragmented approach in the study of IBD has prevailed in the previous decades, highlighting the importance of transitioning towards a more integrated interactome framework.

**Key-words:** knowledge construction; network analysis; experimental research in humans.

Manuscript Doi: 10.1093/ecco-jcc/jjac059

# Graphical_Abstract



Manuscript Doi: 10.1093/ecco-jcc/jjac059

## Introduction

The complexity of inflammatory bowel diseases[1] (IBD) has been recognized in the last three decades[2–7]. The pathogenesis of IBD involves numerous dynamical interactions between components belonging to different functional levels, namely cells, molecules, susceptibility genes, and biological processes and modulating factors, which include the immune-neuro-endocrine system and the intestinal microbiota[1,8–10]. Lately, the notion of the IBD interactome has risen as a conceptual framework to help comprehend the complexity of the disease[11]. In this context, the IBD interactome can be understood as a biological network that includes all relevant components associated with the disease, which are represented as nodes of the network. The putative or demonstrated interactions between components (i.e., nodes) are represented as edges of the network[12]. In this way, by focusing not only on specific components but also on the interactions between them, network representations can provide new and valuable information about pathogenesis, helping to identify key regulatory components[1,12]. Remarkably, the conceptualization of the interactome is quite recent compared to the history of IBD research. Hence, it is reasonable to speculate that the vast knowledge generated on IBD in humans so far, was not obtained completely within this framework. This could have important implications given that network information (nodes and edges) may be missing, or evidence in humans may be lacking, providing a distorted conceptual map of the IBD interactome.

Understanding the interactome is difficult not only due to the complex nature of IBD but also because of the continuous expansion of knowledge regarding its components and their interactions[1,8–10]. It can be expected that for IBD, as for most fields of biomedical research, the number of articles indexed in PubMed/MEDLINE has grown exponentially during the last 40 years, reflecting both technological improvements and new research

Manuscript Doi: 10.1093/ecco-jcc/jjac059

goals[13]. In this scenario, the relative awareness of the medical community regarding the addition of novel components or new findings associated with known components is decreasing[13,14]. More so, scientific discoveries can go unnoticed because they are scattered among different areas of research[15]. This reality, together with a lack of integration of the large amount of existing data from clinical patients, may have contributed to a fragmented or reductionist construction[16] of IBD knowledge over time. Additionally, this may also have hindered the discovery of patterns or anomalies in data that are currently unexplained under existing theories or paradigms[14,17].

Given the enormous volume of biomedical publications and the need to understand IBD by focusing on the interactions between pathophysiological components, the use of text mining tools in combination with network analysis can be useful for the synthesis of knowledge and the detection of patterns[18]. On one hand, by text mining, the information contained in a scientific article can be quantified and arranged in data of interest[18]. For example, the frequency in which specific words (e.g., components of IBD interactome) have been reported, can be gathered. In the past, the text mining of abstracts has supported the processes of knowledge synthesis retrieving important associations between components of complex diseases such as Raynaud's disease[15,19,20] and associations between diseases and microbial pathogens[21]. This kind of associations between pathophysiological components can be graphically represented as networks[22] and characterized with mathematical rigor[23–26]. In that way, networks can help identify the most frequent components cited in the literature as well as the most commonly studied interactions among them. Moreover, by creating a variety of networks using abstracts from different time-periods it is possible to compare how the contributions made to the body of literature change qualitative and quantitatively over time. Specifically, we can assess how the focus of a scientific field may be on particular topics at a

Manuscript Doi: 10.1093/ecco-jcc/jjac059

specific time-point in history, as well as show how certain time periods are more prolific than others. Therefore, the combined approach of text mining and network analysis represents a useful method to explore the literature and understand how knowledge in a field has been constructed, which in general allows to explain the patterns observed in the data.

Our goal was to analyze through abstract's text mining and network analysis how the representativeness of specific components of the different functional levels involved in IBD interactome and their relations have changed over the last 30 years of experimental research in human patients. We aimed to understand how knowledge in this field has been constructed and provide insights on the main challenges in building the IBD interactome.

## Materials and methods

### *General procedure*

Three basic steps were applied for information extraction and analysis as shown in Figure 1 and Table 1. First, a comprehensive bibliographical search in PubMed was performed to retrieve all relevant abstracts of experimental research papers associated with IBD in humans, published between 1990-2020 (available in Supporting Table S1[27]) and then curated (available in Supporting Table S2). Second, abstracts were scanned automatically to detect and quantify key words representative of four functional levels involved in IBD interactome[8–10], namely: cells, molecules, genes and biological processes and modulating factors (Supporting Tables S3-S6). Third, network analysis was used to depict representativeness of these key components and their relations over time and to compare information contained in each decade regarding these components (Table 1). In the networks,

Manuscript Doi: 10.1093/ecco-jcc/jjac059

each node represents a component found in the abstracts of the studied time period. The node size is given by the number of components of the network with which that node is related (co-occurs in abstracts), and the edge (line) width, by the proportion of abstracts in which a given pair of components are referenced together. In this last step, the resulting networks were mathematically characterized using a variety of measures as described in Table 1. Detailed information about each step is presented in Supporting Material 1, Section S1.

### *Software*

MATLAB™ R2018a was used for running all automatic analysis and calculations. We used functions of "wordcloud" in the Text Analytics Toolbox and "graph" function. Custom MATLAB™ code created for additional text preprocessing, curation of the corpus, matrix operations and extracting graph (network) measures, can be found in Supporting Materials 2-3.

## Results

### *General overview of the retrieved abstracts*

A total of 27208 abstracts were originally retrieved in our PubMed search, and 25971 remained after curation. The resulting word cloud (Fig. 2A) showed that the most frequent words appearing in the abstracts were those referring to the disease (i.e., Crohns, UC, CD IBD, colitis). With lower frequencies were those referring to therapy (remission, infliximab), gut localization (colonic), and components from the diverse functional levels involved in IBD such as genes, cells, molecules (cytokines, calprotectin), environmental factors (smoking) and population-related expressions (pediatric). The predominance of these words in abstracts not

only reflects the wide diversity of factors on IBD research in humans but probably the most relevant ones.

The annual publication rate showed a 3-fold exponential increase between 1990 and mid-2020 (Fig. 2B). Journals with scope in IBD and gastroenterology concentrated the majority of the publications (Fig. 2C). In fact, 50% of the publications were distributed among the top 20 out of 2043 journals (Fig. 2C). Meanwhile, journals with more general scopes showed smaller numbers of publications (Fig. 2C). A more detailed analysis revealed that 85 % of the journals contributed at most with 10 publications only (Supporting Material 1, Fig. S4).

***of components of interest in the corpus of abstracts***

Deepening into the analysis of the representativeness of the terms, we focused on four functional levels that have been pointed as key components of human IBD[8–10]: (1) immune-endocrine cells (n=36 terms), (2) molecules (n=135 terms), (3) susceptibility genes (n=985 terms), and (4) biological processes and modulating factors (n=62 terms); for the search, synonyms of these 1218 components of interest were retrieved from catalogs of genes and molecules (for details see Supporting Material, Section S1). The counting frequency for IBD interactome components showed a distribution similar to a power law (Fig. 3A), with few words having high frequencies and most appearing infrequently in the abstracts. The word IBD or its synonyms (Supporting Table S6) was found in almost all abstracts (Fig. 3I). Other words such as generic genetic terminology were present in approximately half of the abstracts (Fig. 3I). However, most of the 1218 terms searched for, although specifically referred as important components in recent review articles[8–10], actually appeared in very few papers. Quantification showed that, even though we considered all component synonyms in our

Manuscript Doi: 10.1093/ecco-jcc/jjac059

search, 82% were mentioned in less than 10 abstracts (Fig. 3A, red dotted line), indicating that the representativeness of these components was less than 0.04 % of the corpus (i.e., <10 out of 25971 abstracts). Consequently, to extract significant relationships, we focused only on components that were found above this threshold. Specifically, 75 % of the cells (27 out of 36), 40 % of the molecules (54 out of 135), 8 % of the genes (80 out of 985) and 85 % of the biological processes and modulating factors (53 out of 62) were found in at least 10 abstracts (Fig. 3B-E, dark blue).

In Figure 3F-I the top most frequent words within each functional level appearing in ≥ 10 abstracts are shown. Frequencies of the full set of components are given in Supporting Table S21. Among the 27 cell types that met the cutoff criteria, the most frequent were generic T cells and T helper 1 cells, epithelial cells, and some blood cells (Fig. 3F). From the 54 molecules retrieved, the most frequent belonged to the "interleukin" class (i.e., IL, IL-6, IL-1, IL-10, IL-8), which appeared in approximately 16% of the abstracts (Fig. 3G). Other interleukins, immunoglobulins, molecules associated with oxidative stress and intestinal integrity were retrieved at a lower frequency. Among the 80 genes fitting the cutoff the most frequent were TNF, NOD2, IFNγ, NF-κB and IFN (Fig. 3H). Regarding the 53 biological processes and modulating factors, the most frequent referred to the disease, generic genetic terminology, *in vivo* studies, time related variables, the microbiota, immune-endocrine pathways or biological processes involved in the disease such as inflammation, barrier integrity, adaptive immunity, as well as references to the adult and child populations (Fig. 3I). References to environmental challenges, origin of the populations studied, other immune-endocrine pathways and *in vitro* techniques were mentioned with a lower frequency.

Manuscript Doi: 10.1093/ecco-jcc/jjac059

To explore potential interactions between components of the IBD interactome, we assessed the co-expression of the most frequent components within each abstract. Specifically, we estimated the level of co-occurrence of the components from the four functional levels described in the previous section (Fig. S5-S8); the corresponding network representations for the whole 30-year period are shown in Figure 4A, C, E and G. It can be seen that the four networks differ not only in the number of components they include (i.e., nodes) but also in their level of connectivity. In networks, lines (i.e., edges) connect components that are expressed together in the same abstracts; note that the width of the line indicates the proportion of abstracts in which a given pair of components are referenced together. Moreover, the size and color of a node is given by the number of connections, or edges incident in it (i.e., node degree). Thus, the network of biological processes and modulating factors showed the highest level of connectivity (Fig. 4G and Table 2), followed by molecular (Fig. 4C) and cellular levels (Fig. 4A).

A more in-depth analysis of the connectivity of the nodes is shown in the frequency histograms of the degree of the nodes (Fig.4 B, D, F and H). With the exception of the gene functional level, more than 50% of the nodes showed a degree over 20 (Fig. 4 B, D, F, H; see to the right of Q2) indicating that these components were frequently cited together. Remarkably, the histogram for the network constructed using the complete set of components showed that ≈20% of them presented a node degree higher than 100 (Fig. 4 I, J), indicating that only a relatively small proportion of the components are cited together. Remarkably, the network constructed using the complete set of components (i.e., complete set) showed that ≈20% of components presented a node degree higher than 100 (Fig. 4 I, J). This high value accounts for the cumulative complexity of the network based on information provided through the three decades. Moreover, this network (Fig. 4I) not only presented a higher

Manuscript Doi: 10.1093/ecco-jcc/jjac059

number of nodes and edges, but also a higher edge entropy (i.e., higher diversity of edges) than any of the functional level networks separately (Table 2). Still, comparatively a decrease in network complexity evaluated as node entropy was found (Table 2), reflecting a lack of information due to absence of relationships between genes and the majority of the other components.

An important feature of network representations shown in Figure 4 is the presence of motifs that are visualized by thick lines connecting a subset of components (these are highlighted in red in Fig. S16D, 18D, 20D, 21D, 23D and 25D). For example, in the network of relationships between cells, a conserved motif was evidenced encompassing cells with a higher individual frequency of occurrence, such as T cells, epithelial cells, neutrophils, red blood cells, macrophages and monocytes (Fig. 3F and supporting material 1, Fig. S16).

***Functional levels: Representativeness of components and their relationships over time***

Network representations for each decade of the four functional levels described in the previous section were also depicted (Supporting Material 1, Section S4, Figs. S15-S25). Quantifications of these networks are shown in Table 2 and Figure 5. We found that essentially the same set of cells has been studied since 1990 (Table 2, number of nodes). The only exceptions are the Natural Killer T cells and the type 3 innate lymphoid cells that were integrated into the network in the second and third decades, respectively (Supporting Material 1, Fig. S15B, C, see arrows). Contrarily, the diversity of molecules showed an increase of ≈45% between the first and second decade (Table 2), in which molecules associated with intestinal epithelium integrity and various interleukins were incorporated into the network (Supporting Material, Fig S17A, B). Coherently, the number of abstracts in which components were co-expressed also increased more pronouncedly between the same period

Manuscript Doi: 10.1093/ecco-jcc/jjac059

(Table 2, number of edges). Unlike the observed with cells, 20% of the nodes became dominant in the network of molecules (Supporting Material 1, Fig. S15 vs S17), namely those of IL-1, IL-6, IL-10, IL-8, IL-2, IL-4, and IgA, IL1R, IL5 and IgG, defining a motif conserved since 1990 (Supporting Material, Fig. S18). Conversely, other molecules, in general, interacted selectively with molecules within their own class but not outside (Supporting Material 1, Figs. S17 and S18). Thus, although information associated to both nodes and edges (Fig. 5 and Table 2, edge and node entropy, respectively) increased over time, there was a conserved pattern of relationships characterized by a dominance of interleukins and "small-range" relationships within classes (Supporting Material 1, Figs. S17 and S18).

The diversity of genes showed a marked 3-fold increase between the first and second decade, and increased at a lower rate thereafter (Table 2, number of nodes, and Supporting Material 1, Fig. S19 A-C). Also, the connectivity increased more pronouncedly between the same time-period, with the number of edges showing a 9-fold and a 1-fold increase by the second and third decade, respectively (Table 2, number of edges, and Supporting Material 1, Fig. S19 A-C). Similar to the network of molecules, 30% of the nodes (28 genes) with high individual frequencies emerged as dominants and defined a motif that was completely visible since the second decade (Supporting Material 1, Fig. S20). Remarkably, a submotif TNF-IFNγ-IFN-IFNα is apparent since 1990 (Supporting Material, Fig. S21). This reduced set of dominant genes concentrated more interactions than other genes (Supporting Material 1, Fig. S19-21). Thus, at this functional level, node-associated information actually decreased over time (Fig. 5B, C, Table 2), with relationships among many genes in the networks absent.

Manuscript Doi: 10.1093/ecco-jcc/jjac059

Finally, biological processes and modulating factors remained almost unchanged over time, with similar number and identity of nodes in the network of each decade (Table 2, Supporting Material 1, Fig. S22 A-C). This indicates that the same set of factors has been studied since 1990, except for some genetic analysis techniques that were integrated to the network in the second decade (i.e., GWAS, QTL, metagenomic). The pattern of relationships shows high-degree nodes dominating these networks (Supporting Material 1, Fig. S22 A-D), defining a well-conserved motif since 1990 (Supporting Material 1, Fig. S23 A-D). Thus, although diverse biological processes and modulating factors as well as their relations have gained relevance over time, slight changes in relationships have occurred with a predominance in the same key factors since 1990.

Although a 1- to 6-fold increase in edge entropy was observed within networks of molecules and genes over time, little or no change in cells and a decrease in biological processes and modulating factors were found (Fig. 5B).

***The network of all components of IBD interactome: evidence that many relations remain unexplored***

Finally, encompassing the information of all networks previously described, we studied the networks of all the 214 components of the IBD interactome obtained for each decade (Supporting Material 1, Fig. S24 A-C). As expected, a critical increase in the diversity of components was observed between the first and second decade, with ≈50% increase in the number of nodes (Table 2). New components continued to be incorporated in the third decade, although this increment was only ≈4% (Table 2). However, the number of edges showed a much larger increment (127%) between the first and second decade and a 46% increase between the second and third decade (Table 2). Moreover, components dominating

Manuscript Doi: 10.1093/ecco-jcc/jjac059

in each functional level network were also dominants in the general network (compare panels A-D in Supplementary Materials 1, Figs. S15, S17, S19, S22, and S24).

As with each functional level, the network of the complete set of components showed an increase in mean connectivity over time (Fig. 5A) with higher connectivity than any of the functional level networks separately. However, node entropy remained constant and lower than in any of the individual functional levels. Moreover, in the last decade a decrease in complexity associated with the entropy of edges was found in the complete set (Fig. 5B) indicating that even when new components (i.e., nodes) are being added to the networks, no new information regarding their relations was incorporated.

**Discussion**

This work presents the first text mining-based analysis that covers the entire spectrum of components of the IBD interactome and the relationships among them. Through our approach we demonstrate that, in general, the number of components studied in the human IBD interactome increased over the last 30 years of research. However, well-conserved patterns of relations among components within and across the functional levels arose. We also shed light on the consistency of the IBD interactome, noting that the vast majority of components (82%) had weak experimental support in humans, showing low individual frequency and/or low co-occurrences or even being completely unexplored in humans. Globally, our findings demonstrate that the IBD interactome conceptualization is contemporary and thus, historical analysis reflects the reductionist approach with which this particular disease has been studied.

Manuscript Doi: 10.1093/ecco-jcc/jjac059

***Fragmented increase of information***

As expected, our analysis on the last 30 years of IBD research in humans revealed that not only the number of publications increased over time (Fig. 3B) but also the information regarding specific components of IBD interactome and their relationships (Table 2). Considering the annual rate of article indexing in PubMed/MEDLINE[28,29] and the explosion of information driven, in part, by technological improvements[13], is reasonable that IBD research has followed this trend as well[30,31]. The general increase in information is also consistent with the remarkable growth and diversification of scientific output and capacity worldwide, which has fueled a distinctive global scientific system, based primarily on research universities, fostered by communication and publishing in English via the Internet, cross-border authorship and mobility of researchers[31].

By analyzing separately each functional level, clear differences on how the information increased over time was observable. Specifically, within the networks of molecules and genes, the number of components and the new information regarding their relationships (entropy of edges) increased over each decade studied (Table 2). This behavior could be attributed to the progress of Human Genome Project from 1990 to 2003[32], as well as the development of high-throughput molecular biology techniques[33–35]. Specifically regarding IBD, the association of specific susceptibility genes as contributing factors became evident in 1999, when pioneering population studies confirmed the increased incidence in relatives (up to the third degree) of the affected individuals[36].

In contrast, within the networks of cells and biological processes and modulating factors, component number and edge entropy almost stabilized in the second decade or even decreased afterwards. This stability over time of the cell networks seems to respond to their

Manuscript Doi: 10.1093/ecco-jcc/jjac059

long-standing scientific history regarding research in general[37]. Contrarily, the stability seen in the biological processes and modulating factor networks, could be related to the heterogeneity of the words included in this functional level and the fact that some of these words represent conceptualizations[38,39]. Even though the impact of milestones regarding cells and biological processes and modulating factors on science is long-standing, new relationships between these components have not been generated for two decades, notoriously without information gain in relation to entropy of edges.

In terms of the network of all components, clearly the body of evidence and information in IBD research in humans has increased over the last 30 years, although singularities inherent to the particular temporal evolution found in each functional level suggest a fragmented accumulation.

***Patterns conserved over time in knowledge construction in human research***

The study of the relationships among components, in contrast to each isolated component, allowed us to identify that the addition of information followed well-defined patterns over time. Specifically, each new node seemed to attach to the existing network according to a "rich-get-richer" rule[40,41], as evidenced by the strengthening over time of both the motifs and the degree of the bigger nodes in each network. That is, IBD components and relations studied with a high frequency in a given decade, were studied even with a higher frequency in the following decades, as shown in other network studies that focus on changes over time[41–47]. For components and relations involved in IBD interactome, this behavior may implicate that the same set of "hypotheses" has been leading research at least during the last 30 years, which is reflected in the relatively small set of components and relations studied more frequently. In all, many aspects of nowadays IBD research in humans seem to be

Manuscript Doi: 10.1093/ecco-jcc/jjac059

reflecting ideas from the early years of these diseases, leading to the well-conserved patterns we found in the networks.

***Potential bias in knowledge construction***

Herein, the topology of the networks essentially indicated that some components and relationships were highly studied over time in contrast to others which have been involved in IBD interactome with a weaker experimental support in humans or not studied at all, an aspect overlooked in most reviews[8–10]. Many genes and molecules that have been theoretically associated with IBD may have been absent in our corpus of abstracts due to a combination of their incipient discovery and the "big data" nature of current molecular high-throughput methods. As these methods allow the analysis of a very huge number of variables at once, individual mention of each one of them in an abstract could be impossible or meaningless (see for example [48]).

The lack of empirical support for many components and relations in the networks observed could also reflect a phenomenon of extrapolation of results from experimental models to humans[49,50]. In this scenario, descriptions presented in review articles and textbooks are composite images reconstituted from data collected from different experimental setups[49]. As a consequence of this epistemic practice, results from validated surrogate models are treated as a legitimate line of evidence. However, at the same time, the overall structure of research also demonstrates that extrapolative inferences should not be considered definitive, but only as partially justified hypotheses subjected to further testing[49]. Thus, based on our analysis, it seems that the majority of the hypothesized components and relations of the human IBD interactome remain to be more thoroughly tested in humans.

Manuscript Doi: 10.1093/ecco-jcc/jjac059

As stated previously, the presence of components and relations highly studied over time may represent empirical support to a given hypothesis. But on the other hand, could be reflecting potential bias in knowledge construction. One source of bias may be related to the exponential growth of scientific publications which makes almost impossible to keep track of all research advances in the scientific literature, leading to fragmented knowledge based on individual specialization[51]. Also relevant is the explanation of how and why scientific ideas change over time or remain stable. It has been claimed that researchers are locked in "thought-collectives" representing community of persons mutually exchanging ideas and maintaining intellectual interaction[16]. Usually there is an agreement on what members of a thought-collective consider evident and what methods are adequate for further research[16].

The asymmetry in the publication of some components over others is dependent both on "thought-collectives" and the specific historical context. For example, Rogler (2013) has pointed that when it was possible to characterize T cells by flow cytometry and surface marker analysis and to discriminate certain subpopulations of T cells, research was focused in autoimmunity and T-cell-mediated factors as the cause of IBD[52]. Also, when innate immune processes were better understood, IBD became an "innate immune disease". More recently, the acquisition of gut bacteria pyrosequencing ability put the entire gut microbiota in the spotlight[52]. To summarize, although technical tools available could make huge improvements in the way the pathophysiology of IBD is approached, also different sources of scientific bias may affect how technological facilities are used. Thus, a reflexive use of information and technologies should guide further research on IBD interactome.

Manuscript Doi: 10.1093/ecco-jcc/jjac059

***Towards (a more integrative) systems biology***

The notorious trend of increasing the number of relations among components of IBD interactome (Table 2, Fig. 6), although biased, supports the contemporary claim for a systems biology approach for understanding IBD pathophysiology. Some good efforts have been recently made in this direction[48,53–57]. However, it is evident that simultaneous evaluation of components from the different functional levels involved is still a challenge, as it is demonstrated for instance by the lack of links among environmental risk challenges (from smoking to the use of macrolides antibiotics) and the rest of the components. In this regard, future perspectives in IBD interactome research should include collaborative studies, phenotyping of patients, assembling integrative bioinformatics and expertise, long-term prospective collections of samples from different -omes relevant to IBD pathogenesis, in order to achieve, through network medicine methods, personalized medicine[1,12]. As it has been discussed, the history and epistemology of IBD research indicates that it is highly probable that new components and relations emerge with future studies.

***Limitations of the study***

One of the potential limitations of our study, added to those already mentioned, is directly related to our methodological approach. Herein, in order to exhibit and represent the known relationships between pathophysiological components of IBD in a visually apprehensible network, we used the analysis of the distribution of frequencies of the components along the corpus of abstracts to extract these relationships. In that way, components studied in at least 0.04% of the abstracts were taken into account given that they were expected, by probability, to have significant relationships with other components. It is feasible that with this methodological approach we excluded from the networks some key and

underrepresented relationships between components studied in very few articles, but whose importance for the field and for the understanding of the disease are crucial. Therefore, our synthesis of knowledge construction in itself can be incurring in a secondary bias towards components of high abundance in the literature. It is worth noting that in the field of interactomics, particularly at the molecular level, it is widely recognized that since the technique used determines the type of relationships found, the procedures are always prone to bias[58–60]. Thus, a future synthesis could implement other approaches to filter components and shed light on different aspects of knowledge construction than those studied here.

Another limitation of our synthesis is related with the epistemology of the interactome in itself and the evolving nature of knowledge construction. Considering that the interactome framework focuses on interactions between components and that virtually infinite components can be added over time, it is reasonable to ask how expandable interactomes might be and how reliable are the interactomes built with the information we have so far. With this in mind, it could be said that our synthesis is as limited and reductionist as science itself. In this regard, interactomes in general are not complete with perhaps some exceptions[61–63], which is basically associated with the fact that methodologies still need to be improved and that, in the end, interactomes are snapshots of ongoing scientific efforts[60].

An additional limitation of our analysis is that with the synonyms used to scan for "time related variables" we couldn´t properly account for the low frequency of studies that actually focus on the temporal dynamics of the disease. As has been demonstrated in many fields of biology and medicine[64–72], the role of time both at short- and large-scales seems to be key to understand complex biological responses and diseases. It is likely that the dynamics

Manuscript Doi: 10.1093/ecco-jcc/jjac059

of onset and relapse in IBD as well as the propensity to manifest the disease later in life, for mentioning only two examples, do not fall outside this umbrella.

Finally, another potential source of bias in our synthesis may result from the fact that we have not filtered the components to distinguish articles with conflicts of interest as funding sources. However, since the database we built is easily accessible, it can be studied further to explore this and similar issues. For example, networks built from articles using commercial compounds can be compared to networks built from articles studying only generic compounds, thus accounting for the impact of funding agencies in knowledge construction.

**Concluding remarks**

Even recognizing the limitations of our analysis, it is important to highlight that this is the first study in analyzing IBD interactome research in humans using text-mining, information theory and network analysis methods. Our approach allowed us to analyze the huge amount of specific information that has been gathered in the last 30 years. Nevertheless, the main challenge to understand this complex disease may be to recognize that knowledge construction in the field is still characterized by reductionism and fragmentation.

Manuscript Doi: 10.1093/ecco-jcc/jjac059

**Supporting information**

Supporting Material 1 includes: Extended Materials and Methods; Section 1 Bibliographical search (Supporting Figs. S1-4); Section S2 Matrices of co-occurrence (Supporting Figs. S5-9); Section S3 Matrices of conditional co-occurrence (Supporting Figs. S10-14); Section S4 Networks (Supporting Figs. S15-25); Section S5 Changes over time in biological processes and modulating factors (Supporting Fig. S26).

Supporting Materials 2-3 containing MATLAB™ custom code and Supporting Tables S1-S21 have been uploaded to Figshare (private link for ease of access during reviewing process: https://figshare.com/s/9ec3586d516bcac8cd7c) as follows:

Supporting Material S2_Text preprocessing, curation and automatic scanning of abstracts.

Supporting Material S3_ Matrix operations and graph measures.

Supporting Table S1_Raw corpus of abstracts; Supporting Table S2_Curated corpus of abstracts; Supporting Table S3_Target Matrix of Cells; Supporting Table S4_Target Matrix of Molecules; Supporting Table S5_Target Matrix of Genes; Supporting Table S6_Target Matrix of Biological Processes and Modulating Factors; Supporting Table S7_Summary matrix for the complete set of components; Supporting Table S8_Subset summary matrix for the complete set of components and the whole time-period; Supporting Table S9_Subset summary matrix for the complete set of components and the first decade; Supporting Table S10_Subset summary matrix for the complete set of components and the second decade; Supporting Table S11_Subset summary matrix for the complete set of components and the third decade; Supporting Table S12_Matrix of co-occurrence for the complete set of components and the whole time-period; Supporting Table S13_Matrix of co-occurrence for the complete set of components and the first decade; Supporting Table S14_Matrix of co-occurrence for the complete set of components and the second decade; Supporting Table

Manuscript Doi: 10.1093/ecco-jcc/jjac059

S15_Matrix of co-occurrence for the complete set of components and the third decade; Supporting Table S16_Class definition within cells' functional level; Supporting Table S17_Class definition within molecules' functional level; Supporting Table S18_Class definition within genes' functional level; Supporting Table S19_Class definition within biological processes and modulating factors' functional level; Supporting Table S20_Journals; Supporting Table S21_Individual accumulated frequency.

## Conflict of Interest

The authors have no conflict of interest to declare.

## Authors' contributions

**M.E.F.:** Conceptualization, Data Collection, Data Curation, Code Writing, Data Analysis, Visualization, Data Interpretation, Writing -original draft.

**F.N.N.:** Conceptualization, Data Interpretation, Writing - Review & Editing.

**L.B.M.:** Data Collection, Data Curation, Writing - Review & Editing.

**C.E.J:** Data Collection, Data Curation, Writing - Review & Editing.

**J.M.K.:** Conceptualization, Code Writing, Data Analysis, Visualization, Data Interpretation, Funding acquisition, Resources, Supervision, Writing - Review & Editing.

**S.G.C:** Conceptualization, Data Interpretation, Project Administration, Supervision, Writing - Review & Editing.

All authors have approved the final version of the manuscript for publication.

Manuscript Doi: 10.1093/ecco-jcc/jjac059

**Funding**

This research was supported by grants from FONCyT (Fondo para la Investigación Científica y Tecnológica, Préstamos BID-PICT-2016-0282; 2018-3398). SGC, JMK, FNN are career members of CONICET. MEF and LBM, and CEJ have a post-doctoral and doctoral scholarship from the later institution, respectively.

**Data availability**

This unique and curated database is freely available[27] and can be a useful tool for researchers to pinpoint key publications studying specific components or relationships, as well as for further epistemological studies.

All raw data and custom code for analysis underlying this article has been uploaded to Figshare, the DOI https://doi.org/10.6084/m9.figshare.16906534 has been reserved and will be publicly available upon acceptance of manuscript. A Private link have been cited in manuscript for ease of access during reviewing process: https://figshare.com/s/9ec3586d516bcac8cd7c.

Manuscript Doi: 10.1093/ecco-jcc/jjac059

## References


1 De Souza HSP, Fiocchi C, Iliopoulos D. The IBD interactome: An integrated view of aetiology, pathogenesis and therapy. *Nat Rev Gastroenterol Hepatol* 2017;**14**:739–49. https://doi.org/10.1038/nrgastro.2017.110.

2 Satsangi J, Parkes M, Louis E, Hashimoto L, Kato N, Welsh K, Terwilliger JD, Lathrop GM, Bell JI, Jewell DP. Two stage genome-wide search in inflammatory bowel disease provides evidence for susceptibility loci on chromosomes 3, 7 and 12. *Nat Genet* 1996;**14**:199–202. https://doi.org/10.1038/ng1096-199.

3 Parkes M, Satsangi J, Lathrop GM, Bell JI, Jewell DP. Susceptibility loci in inflammatory bowel disease. *Lancet* 1996;**348**:1588. https://doi.org/10.1016/S0140-6736(05)66204-6.

4 Gaya D, Russell R, Nimmo E, Satsangi K. New genes in IBD, lessons from complex diseases. *Lancet* 2006;**367**:1271–84.

5 Mathew CG, Lewis CM. Genetics of inflammatory bowel disease: Progress and prospects. *Hum Mol Genet* 2004;**13**:161–8. https://doi.org/10.1093/hmg/ddh079.

6 Cleynen I, Halfvarsson J. How to approach understanding complex trait genetics-inflammatory bowel disease as a model complex trait. *United Eur Gastroenterol J* 2019;**7**:1426–30. https://doi.org/10.1177/2050640619891120.

7 Ek WE, D'Amato M, Halfvarson J. The history of genetics in inflammatory bowel disease. *Ann Gastroenterol* 2014;**27**:294–303.

8 Graham DB, Xavier RJ. Pathway paradigms revealed from the genetics of inflammatory bowel disease. *Nature* 2020;**578**:527–39.

https://doi.org/10.1038/s41586-020-2025-2.

9 Vennou KE, Piovani D, Kontou PI, Bonovas S, Bagos PG. Methods for multiple outcome meta-analysis of gene-expression data. *MethodsX* 2020;**7**:100834. https://doi.org/10.1016/j.mex.2020.100834.

10 Lahue KG, Lara MK, Linton AA, Lavoie B, Fang Q, McGill MM, Crothers JW, Teuscher C, Mawe GM, Tyler AL, Mahoney JM, Krementsov DN. Identification of novel loci controlling inflammatory bowel disease susceptibility utilizing the genetic diversity of wild-derived mice. *Genes Immun* 2020. https://doi.org/10.1038/s41435-020-00110-8.

11 De Souza H, Fiocchi C. Immunopathogenesis of IBD: current state of the art. *Nat Rev Gastroenterol Hepatol* 2016;**13**:13–27.

12 De Souza HSP, Fiocchi C. Network Medicine: A Mandatory Next Step for Inflammatory Bowel Disease. *Inflamm Bowel Dis* 2018;**24**:671–9. https://doi.org/10.1093/ibd/izx111.

13 Wren JD, Garner HR. Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. *Bioinformatics* 2004;**20**:191–8. https://doi.org/10.1093/bioinformatics/btg390.

14 Wren JD, Bekeredjian R, Stewart JA, Shohet R V., Garner HR. Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics* 2004;**20**:389–98. https://doi.org/10.1093/bioinformatics/btg421.

15 Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med* 1986;**30**:7–18. https://doi.org/10.1353/pbm.1986.0087.

16 Fleck L. *Genesis and Development of a Scientific Fact*. 1st edn. London: The University of Chicago Press; 1935.

17 Koshland DE. The Cha-Cha-Cha theory of scientific discovery. *Science (80- )* 2007;**317**:761–2. https://doi.org/10.1126/science.1147166.

18 Thilakaratne M, Falkner K, Atapattu T. A systematic review on literature-based discovery workflow. *PeerJ Comput Sci* 2019;**5**:e235.

19 DiGiacomo RA, Kremer JM, Shah DM. Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study. *Am J Med* 1989;**86**:158–64. https://doi.org/10.1016/0002-9343(89)90261-1.

20 Smalheiser NR, Swanson DR. Using ARROWSMITH a computer-assisted approach toformulating and assessing scientific hypotheses. *Comput Methods Programs Biomed* 1998;**57**:149–53. https://doi.org/https://doi.org/10.1016/S0169-2607(98)00033-9.

21 Sintchenko V, Anthony S, Phan XH, Lin F, Coiera EW. A PubMed-wide associational study of infectious diseases. *PLoS One* 2010;**5**:. https://doi.org/10.1371/journal.pone.0009535.

22 Xun G, Jha K, Gopalakrishnan V, Li Y, Zhang A. Generating medical hypotheses based on evolutionary medical concepts. *Proc - IEEE Int Conf Data Mining, ICDM* 2017:535–44. https://doi.org/10.1109/ICDM.2017.63.

23 Williams G. *Chaos theory tamed*. Washington, D.C: Joseph Henry Press; 1997.

24 Zenil H, Kiani NA, Tegnér J. A review of graph and network complexity from an algorithmic information perspective. *Entropy* 2018;**20**:1–15. https://doi.org/10.3390/e20080551.

25 Bianconi G. The entropy of randomized network ensembles. *Europhys Lett* 2007;**81**:28005. https://doi.org/10.1209/0295-5075/81/28005.

26 Bassett DS, Sporns O. Network neuroscience. *Nat Neurosci* 2017;**20**:353–64. https://doi.org/10.1038/nn.4502.Network.

27 [dataset] Fernandez, Maria E.; Nazar, Franco N.; Moine, Luciana B.; Jaime, Cristian E.; Kembro, Jackelyn M.; Correa, Silvia G. (2021). Inflammatory Bowel Disease (IBD) Interactome: text database and analyzed data of experimental research in humans between 1990-2020, Figshare, Dataset, https://doi.org/10.6084/m9.figshare.16906534

28 Lu Z. PubMed and beyond: A survey of web tools for searching biomedical literature. *Database* 2011;**2011**:1–13. https://doi.org/10.1093/database/baq036.

29 Wilde M De. *From Information Extraction to Knowledge Discovery: Semantic Enrichment of Multilingual Content with Linked Open Data*. Université libre de Bruxelles; 2016 (thesis).

30 Fontelo P, Liu F. A review of recent publication trends from top publishing countries. *Syst Rev* 2018;**7**:1–9. https://doi.org/10.1186/s13643-018-0819-1.

31 Marginson S. What drives global science? The four competing narratives. *Stud High Educ* 2021:1–19. https://doi.org/10.1080/03075079.2021.1942822.

32 *Human Genome Project Timeline of Events*. Bethesda, MD. National Human Genome Research Institute. 2019. https://www.genome.gov/human-genome-project/Timeline-of-Events (Accessed July 8, 2021).

33 Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome

sequencing. *J Appl Genet* 2011;**52**:413–35. https://doi.org/10.1007/s13353-011-0057-x.

34 Soon WW, Hariharan M, Snyder MP. High-throughput sequencing for biology and medicine. *Mol Syst Biol* 2013;**9**:1–14. https://doi.org/10.1038/msb.2012.61.

35 Darwish IA. Immunoassay Methods and their Applications in Pharmaceutical Analysis: Basic Methodology and Recent Advances. *Int J Biomed Sci* 2006;**2**:217–35.

36 Orholm M, Fonager K, Sørensen HT. Risk of ulcerative colitis and Crohn's disease among offspring of patients with chronic inflammatory bowel disease. *Am J Gastroenterol* 1999;**94**:3236–8. https://doi.org/10.1111/j.1572-0241.1999.01526.x.

37 Hook R. *Micrographia: or, Some physiological descriptions of minute bodies made by magnifying glasses. With observations and inquiries thereupon*. 1st edn. London: Jo. Martyn and Ja. Allestry Printers to the Royal Society; 1665.

38 Nuyts J. *Aspects of a cognitive-pragmatic theory of language: on cognition, functionalism, and grammar*. 1st edn. Amsterdam: John Benjamins Publishing Company; 1992.

39 Wittgenstein L, Anscombe G. *Philosophical Investigations*. 6th edn. Oxford: Blackwell; 1997.

40 Leskovec J, Kleinberg J, Faloutsos C. Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. *KDD '05 Proc Elev ACM SIGKDD Int Conf Knowl Discov Data Min* 2005:177–187.

41 Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the internet topology. *ACM SIGCOMM Comput Commun Rev* 1999;**29**:251–62.

42 Katz J. Scale-Independent Bibliometric Indicators. *Measurement* 2009;**3**:24–8. https://doi.org/10.1207/s15366359mea0301.

43 Redner S. Citation Statistics From More Than a Century of Physical Review. *ArXiv:Physics/0407137* 2004:1–12.

44 Bi Z, Faloutsos C, Korn F. The 'DGX' distribution for mining massive, skewed data. *Proc Seventh ACM SIGKDD Int Conf Knowl Discov Data Min* 2001:17–26. https://doi.org/10.1145/502512.502521.

45 Barabási A-L, Albert R. Emergence of Scaling in Random Networks. *Science (80- )* 1999;**286**:509–12.

46 Huberman B, Adamic L. Growth dynamics of the World-Wide Web. *Nature* 1999;**401**:131.

47 Kumar R, Raghavan P, Rajagopalan S, Tomkins A. Trawling the Web for emerging cyber-communities. *Comput Networks* 1999;**31**:1481–93. https://doi.org/10.1016/S1389-1286(99)00040-7.

48 Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, *et al*. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 2019;**569**:655–62. https://doi.org/10.1038/s41586-019-1237-9.

49 Baetu TM. The 'big picture': The problem of extrapolation in basic research. *Br J Philos Sci* 2016;**67**:941–64. https://doi.org/10.1093/bjps/axv018.

50 Saeidnia S, Manayi A, Abdollahi M. From in vitro Experiments to in vivo and Clinical Studies; Pros and Cons. *Curr Drug Discov Technol* 2015;**12**:218–24. https://doi.org/10.2174/1570163813666160114093140.

51 Cheadle C, Cao H, Kalinin A, Hodgkinson J. Advanced literature analysis in a Big Data world. *Ann N Y Acad Sci* 2017;**1387**:25–33. https://doi.org/10.1111/nyas.13270.

52 Rogler G. The history and philosophy of inflammatory bowel disease. *Dig Dis* 2013;**31**:270–7. https://doi.org/10.1159/000354676.

53 Borren NZ, Plichta D, Joshi AD, Bonilla G, Sadreyev R, Vlamakis H, Xavier RJ, Ananthakrishnan AN. Multi-"-Omics" Profiling in Patients With Quiescent Inflammatory Bowel Disease Identifies Biomarkers Predicting Relapse. *Inflamm Bowel Dis* 2020;**26**:1524–32. https://doi.org/10.1093/ibd/izaa183.

54 Howell KJ, Kraiczy J, Nayak KM, Gasparetto M, Ross A, Lee C, Mak TN, Koo BK, Kumar N, Lawley T, Sinha A, Rosenstiel P, Heuschkel R, Stegle O, Zilbauer M. DNA Methylation and Transcription Patterns in Intestinal Epithelial Cells From Pediatric Patients With Inflammatory Bowel Diseases Differentiate Disease Subtypes and Associate With Outcome. *Gastroenterology* 2018;**154**:585–98. https://doi.org/10.1053/j.gastro.2017.10.007.

55 Douglas GM, Hansen R, Jones CMA, Dunn KA, Comeau AM, Bielawski JP, Tayler R, El-Omar EM, Russell RK, Hold GL, Langille MGI, Van Limbergen J. Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. *Microbiome* 2018;**6**:13. https://doi.org/10.1186/s40168-018-0398-3.

56 Titz B, Gadaleta RM, Lo Sasso G, Elamin A, Ekroos K, Ivanov NV, Peitsch MC, Hoeng J. Proteomics and Lipidomics in Inflammatory Bowel Disease Research: From Mechanistic Insights to Biomarker Identification. *Int J Mol Sci* 2018;**19**:. https://doi.org/10.3390/ijms19092775.

57 Jin L, Li L, Hu C, Paez-Cortez J, Bi Y, Macoritto M, Cao S, Tian Y. Integrative Analysis of Transcriptomic and Proteomic Profiling in Inflammatory Bowel Disease Colon Biopsies. *Inflamm Bowel Dis* 2019;**25**:1906–18. https://doi.org/10.1093/ibd/izz111.

58 Futschik ME, Chaurasia G, Herzel H. Comparison of human protein-protein interaction maps. *Bioinformatics* 2007;**23**:605–11. https://doi.org/10.1093/bioinformatics/btl683.

59 Kiemer L, Cesareni G. Comparative interactomics: comparing apples and pears? *Trends Biotechnol* 2007;**25**:448–54. https://doi.org/10.1016/j.tibtech.2007.08.002.

60 von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 2002;**417**:399–403. https://doi.org/10.1038/nature750.

61 Kamburov A, Herwig R. ConsensusPathDB 2022: molecular interactions update as a resource for network biology. *Nucleic Acids Res* 2022;**50**:D587–95. https://doi.org/10.1093/nar/gkab1128.

62 Karagoz K, Arga KY. Assessment of high-confidence protein-protein interactome in yeast. *Comput Biol Chem* 2013;**45**:1–8. https://doi.org/10.1016/j.compbiolchem.2013.03.002.

63 Pesch R, Zimmer R. Complementing the Eukaryotic Protein Interactome. *PLoS One* 2013;**8**:e66635. https://doi.org/10.1371/journal.pone.0066635.

64 Aon MA, Cortassa S. *Dynamic Biological Organization: fundamentals as applied to cellular systems*. 1st edn. United Kingdom: Chapman and Hall; 1997.

65 Goldberger AL, Amaral LAN, Hausdorff JM, Ivanov PC, Peng CK, Stanley HE. Fractal dynamics in physiology: Alterations with disease and aging. *Proc Natl Acad Sci U S A* 2002;**99**:2466–72.

66 Fernandez ME, Labaque MC, Orso G, Marin RH, Kembro JM. Short- and long-term dynamics of the physiological and behavioral response to heat stress and thymol supplementation in Japanese quail. *J Therm Biol* 2021;**97**:102876. https://doi.org/10.1016/j.jtherbio.2021.102876.

67 Guzmán DA, Flesia AG, Aon MA, Pellegrini S, Marin RH, Kembro JM. The fractal organization of ultradian rhythms in avian behavior. *Sci Rep* 2017;**7**:684. https://doi.org/10.1038/s41598-017-00743-2.

68 Ivanov PC, Hu K, Hilton MF, Shea SA, Stanley HE. Endogenous circadian rhythm in human motor activity uncoupled from circadian influences on cardiac dynamics. *Proc Natl Acad Sci* 2007;**104**:20702–7. https://doi.org/10.1073/pnas.0709957104.

69 Aon MA, Roussel MR, Cortassa S, O'Rourke B, Murray DB, Beckmann M, Lloyd D. The Scale-Free Dynamics of Eukaryotic Cells. *PLoS One* 2008;**3**:e3624. https://doi.org/10.1371/journal.pone.0003624.

70 Lloyd D, Aon MA, Cortassa S. Why Homeodynamics, Not Homeostasis? *Sci World J* 2001;**1**:133–45. https://doi.org/10.1100/tsw.2001.20.

71 Lin A, Liu KKL, Bartsch RP, Ivanov PC. Dynamic network interactions among distinct brain rhythms as a hallmark of physiologic state and function. *Commun Biol* 2020;**3**:197. https://doi.org/10.1038/s42003-020-0878-4.

Manuscript Doi: 10.1093/ecco-jcc/jjac059

72 Cortassa S, Aon M, editors. *Computational Systems Biology in Medicine and Biotechnology: Methods and Protocols*. 1st edn. Springer US; 2022.

Manuscript Doi: 10.1093/ecco-jcc/jjac059

## Figure Legends

**Fig. 1. Information-extraction procedure.** First, a bibliographic search in PubMed was conducted to retrieve experimental research on IBD in humans between the years 1990-2020. Second, for each decade, the text of each abstract was automatically scanned to search for 1218 specific components of the multiple functional levels involved in IBD interactome according to recent reviews. In the Figure, molecules are exemplified as a functional level in the target matrix. For each component, semantic synonyms were recovered from catalogs and domain-knowledge. With the occurrence of at least one mention of the component in any of its synonyms in each abstract a summary matrix was constructed. Third, with the accumulated frequency of occurrence of each component along the corpus of abstracts a matrix of co-occurrence was constructed. The matrix of co-occurrence was depicted as a network, where each component is a node and the co-occurrence of two components in the same abstract (i.e., a relation) is represented as an edge. The node size is given by its degree, i.e., the number of edges incident on it. The edge thickness is given by the frequency of co-occurrence between two components.

**Fig. 2. General overview of the retrieved abstracts.** (A) Word cloud depicting the top 120 most frequent words in the corpus of abstracts after removing stop words such as “the”, “a”, “among”, etc.; (B) Number of papers published per year between 1990 and 2020 in the context of IBD in humans, obtained as a result of our search in PubMed. Solid and dotted red lines denote exponential fitting and prediction bounds at 95%, respectively ($y = 311{,}92e^{0{,}057x}$; $R^2 = 0.96$; curve was fitted between 1990 and 2019); (C) Ranking of top 60 journals (of a total of 2043 journals) in which articles of our search in PubMed were published. Grey dashed lines denote quartiles of 25 and 50% of the distribution of all abstracts published

(complete set of the 2043 journals as well as their complete names are given in Supporting Table S20).

**Fig. 3. of IBD interactome components along abstracts for the whole 30-year period.** (A) Distribution histograms of all components of interest (immune-endocrine cells, molecules, susceptibility genes, biological processes and modulating factors) in our corpus of abstracts. Red dashed lines denote the cutoff criteria of frequency ≥ 10 along the corpus of abstracts. Percentage of components of interest that were not found or that showed an absolute accumulated frequency lower than 10 or higher or equal to 10, within the functional level of (B) cells, (C) molecules, (D) genes and, (E) biological processes and modulating factors. The absolute frequency resulting of the summation of frequencies for a given component including all synonyms for (F) immune-endocrine cells, (G) the top 50 molecules, (H) the top 50 susceptibility genes and (I) all biological processes and modulating factors.

**Fig. 4. Network connectivity for each functional level and for the complete set of components considering the whole time-period between 1990-2020.** Network representation and histograms of distribution (%) of the degree of their nodes for (A, B) cells, (C, D) molecules, (E, F) genes, (G, H) biological processes and environmental factors, and (I, J) the complete set of components for the accumulated period 1990-2020. Nodes represent the different components, and edges the co-occurrence of a pair of components. The size and color of a node, is given by the number of edges incident in it (node degree), i.e., components with which it is related, ranging from small-sized and blue (lowest degree) to big-sized and red (highest degree). The width of an edge is given by its weight (i.e., the frequency of co-occurrence of a given pair of components), being the thicker edges those of higher weight. Incomplete light brown outer circles identify different classes of components in the network

Manuscript Doi: 10.1093/ecco-jcc/jjac059

of each functional level or the functional level in the network of all components. Q1, Q2, Q3, Q4 are the first, second, third and fourth quartile of the distribution.

**Fig. 5. Temporal dynamics of the information contained in the articles along decades: comparison among functional levels and the complete set of components.** (A) Mean connectivity, (B) edges entropy and (C) nodes entropy of the networks from each functional level and for the complete set of components over decades. Mean connectivity was calculated as the average number of edges per node. Both edge and node entropies (measures of complexity) were calculated using the formula of Shannon's entropy (see details Supplementary Section 1, Section S1). For edge entropy the probability that two components co-occur with a given weight/in a given frequency was used (diversity of co-occurrence od components); for node entropy, the relative frequency in which a given component/word was found in the abstracts analyzed was used (relative frequency of components).

Manuscript Doi: 10.1093/ecco-jcc/jjac059

## Tables

**Table 1.** Information-extraction and analysis procedure

| ***Step 1: Bibliographical search to retrieve relevant abstracts*** |
|---|
| A PubMed query was constructed to retrieve experimental research on IBD in humans, published between July 1990 and June 2020. Metadata associated with articles was managed using Zotero 5.0 for Windows®. |
| ***Step 2: Automatic scanning of abstracts to detect components of the IBD interactome*** |
| **a) Database curation and text preprocessing:** studies performed only in animal models were excluded from the database (curation). Abstracts' text was prepared to allow the identification and quantification of specific words contained in them. MATLAB™ R2018a was used in this step. |
| **b) Word cloud representation** |
| **c) Design of *target matrices* of components of interest representative of four functional levels pointed as key components of human IBD**: (i) immune-endocrine cells (n=36), (ii) molecules (n=135), (iii) susceptibility genes (n=985), and (iv) biological processes and modulating factors (n=62). |
| **d) Automatic detection of the components included in the *target matrices* along the abstracts and obtention of the frequency of occurrence of each component (*summary matrix*)** |
| **e) Filtering out infrequent components with individual accumulated frequencies ≤ 10 along the abstracts.** |
| **f) Splitting the corpus of abstracts into subsets associated with time period (decades) and functional level.** |
| **g) Calculating the frequencies of co-occurrence (*matrix of co-occurrence*) between components that are co-cited within the same abstract, for each functional level and each decade.** |
| ***Step 3: Network analysis for representation of the state of knowledge over time*** |
| **a) Network construction:** the components of the IBD interactome were represented as nodes and the frequency of co-occurrence between two components as edges. Networks were characterized through several parameters:<br>*Number of nodes and edges*: accounts for the diversity of components and their relations.<br>*Degree of a node*: accounts for the number of components with which a node is related.<br>*Entropy of edges and nodes*: accounts for the information content and complexity of the network. |

Manuscript Doi: 10.1093/ecco-jcc/jjac059

*Average weight of edges*: accounts for the frequency of co-occurrence between components of the network.

*Mean connectivity*: accounts for the extent to which all components of the network are connected, independently of the frequency with which they are studied together.

**b) Network motifs:** motifs are constituted by those edges that present a weight higher than the average weight of all edges of the network.

Manuscript Doi: 10.1093/ecco-jcc/jjac059

**Table 2.** Descriptive measurements of the networks for each functional level (cells, molecules, genes and biological processes and modulating factors) and the complete set of components over decades.

| *Functional level* | *Time Period** | Number of Nodes | Number of Edges | Edge weight | Edge entropy | Node entropy | Mean connectivity |
|---|---|---|---|---|---|---|---|
| Cells | 1990-1999 | 25 | 142 | 588.03 | 3.99 | 0.84 | 5.68 |
| | 2000-2009 | 26 | 201 | 943.78 | 4.29 | 0.85 | 7.73 |
| | 2010-2020 | 27 | 229 | 1522.27 | 4.29 | 0.85 | 8.48 |
| | 1990-2020 | 27 | 259 | 2400.80 | 4.83 | 0.86 | 9.59 |
| Molecules | 1990-1999 | 35 | 139 | 591.37 | 2.21 | 0.72 | 3.97 |
| | 2000-2009 | 51 | 340 | 654.12 | 3.81 | 0.71 | 6.67 |
| | 2010-2020 | 54 | 530 | 783.21 | 4.93 | 0.73 | 9.81 |
| | 1990-2020 | 54 | 611 | 1177.90 | 4.92 | 1.03 | 11.31 |
| Genes | 1990-1999 | 18 | 26 | 319.23 | 0.71 | 0.70 | 1.44 |
| | 2000-2009 | 68 | 277 | 322.74 | 3.58 | 0.68 | 4.07 |
| | 2010-2020 | 80 | 566 | 295.41 | 5.03 | 0.66 | 7.08 |
| | 1990-2020 | 80 | 649 | 408.17 | 4.76 | 0.68 | 8.11 |
| Biological processes and modulating factors | 1990-1999 | 48 | 540 | 4217.04 | 4.12 | 0.61 | 11.25 |
| | 2000-2009 | 53 | 846 | 6765.25 | 5.31 | 0.64 | 15.96 |
| | 2010-2020 | 53 | 1001 | 11257.24 | 5.13 | 0.66 | 18.89 |
| | 1990-2020 | 53 | 1065 | 18093.00 | 5.07 | 0.65 | 20.09 |
| Complete set | 1990-1999 | 135 | 2571 | 1529.09 | 7.60 | 0.60 | 19.04 |
| | 2000-2009 | 205 | 5853 | 1825.44 | 10.47 | 0.60 | 28.55 |
| | 2010-2020 | 214 | 8573 | 2349.98 | 9.47 | 0.60 | 40.06 |
| | 1990-2020 | 214 | 9639 | 3606.40 | 9.70 | 0.59 | 45.04 |

*Articles retrieved between 1990-1999 (4327); between 2000-2009 (7567); between 2010-2020 (14077); total between 1990-2020 (25971).

Manuscript Doi: 10.1093/ecco-jcc/jjac059

## Figures

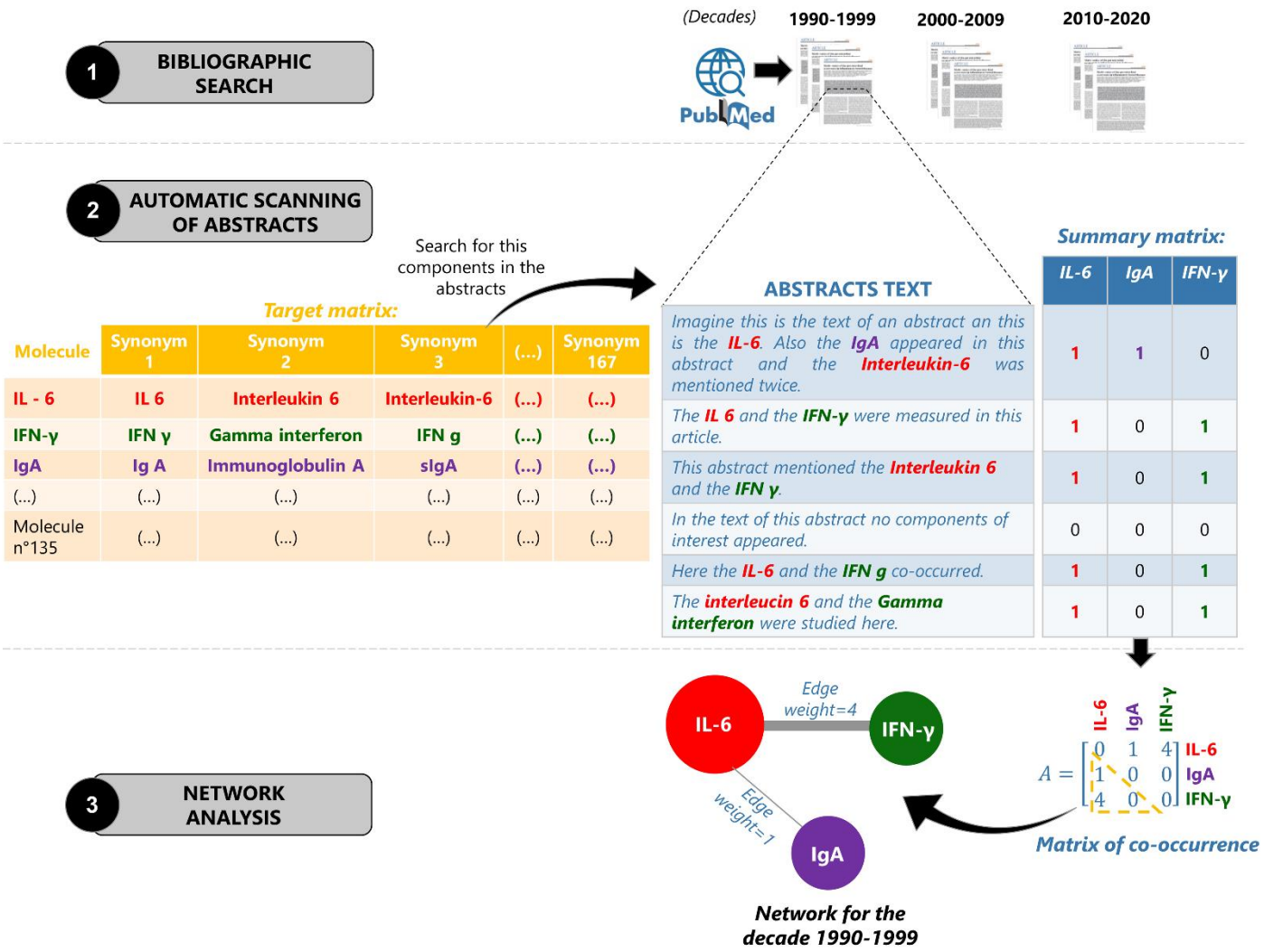



**Fig. 1. Information-extraction procedure.**

Manuscript Doi: 10.1093/ecco-jcc/jjac059

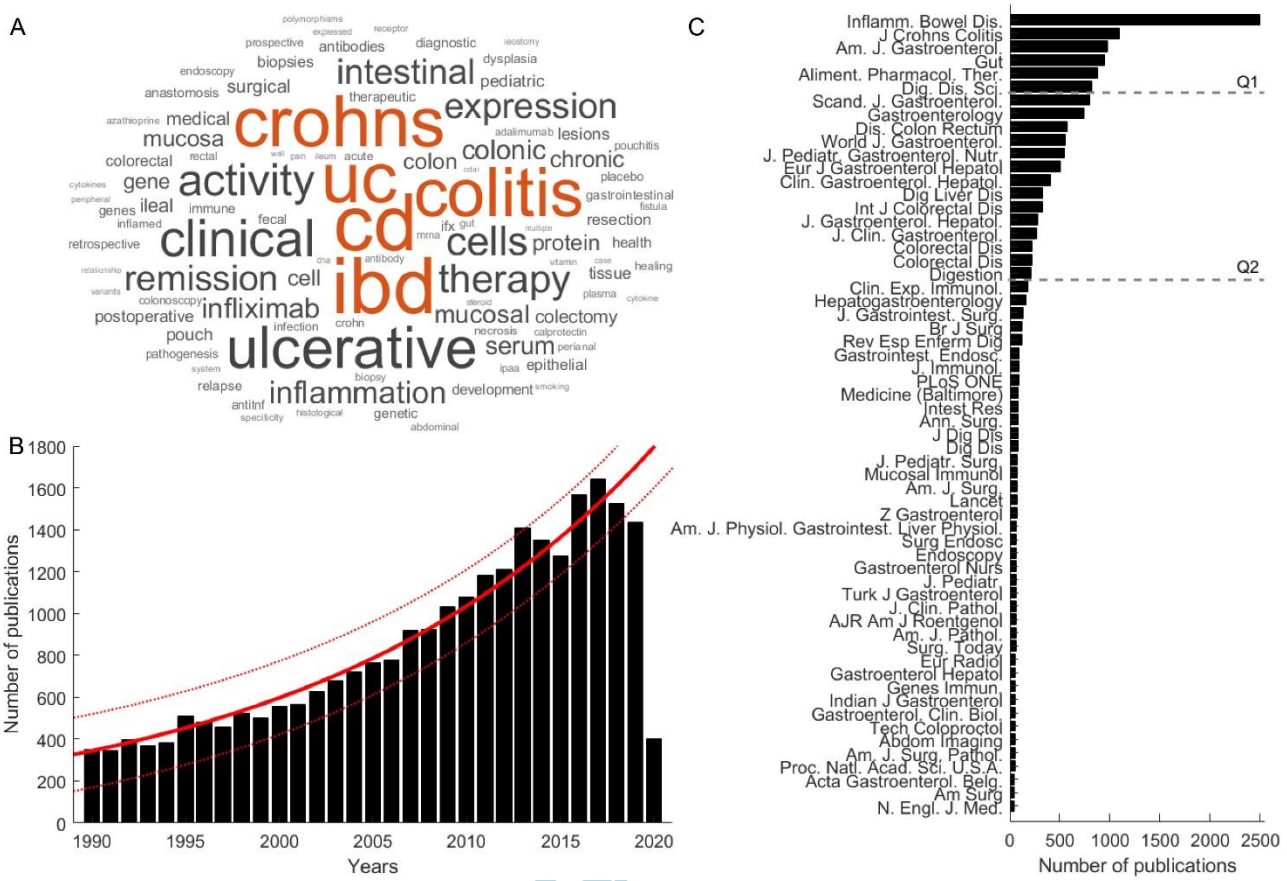



**Fig. 2. General overview of the retrieved abstracts.**

Manuscript Doi: 10.1093/ecco-jcc/jjac059

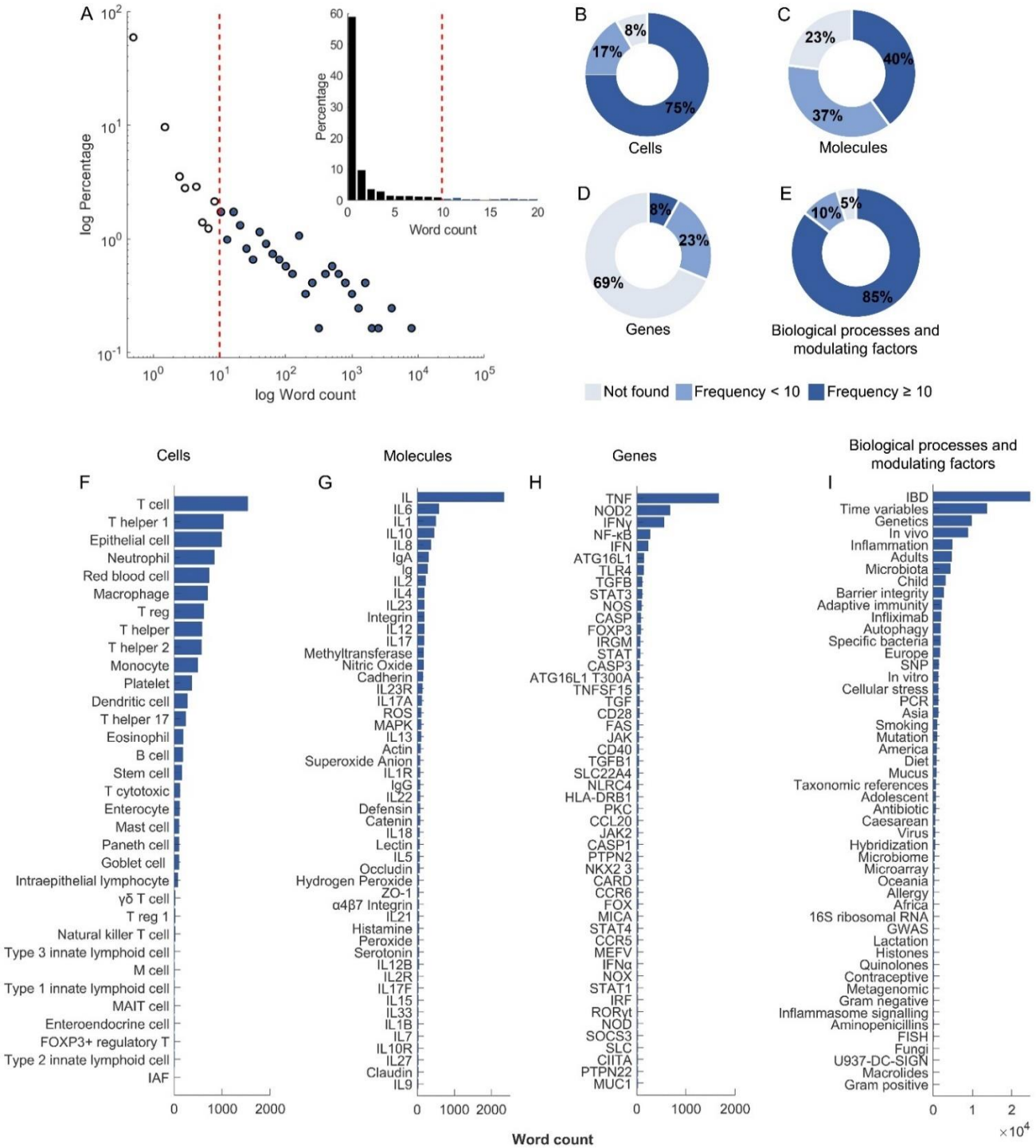



**Fig. 3. of IBD interactome components along abstracts for the whole 30-year period.**

Manuscript Doi: 10.1093/ecco-jcc/jjac059

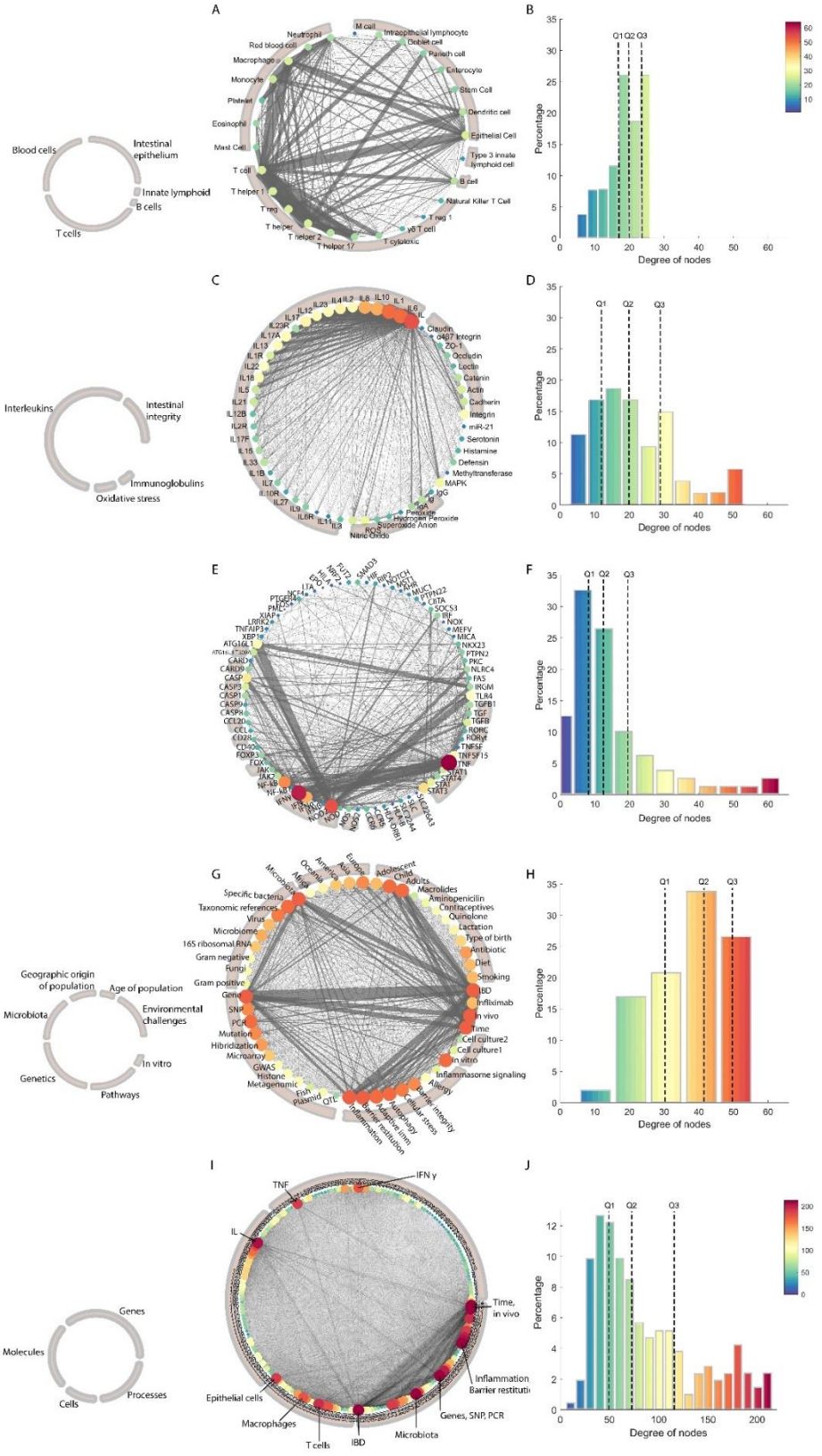



**Fig. 4. Network connectivity for each functional level and for the complete set of components considering the whole time-period between 1990-2020.**

Manuscript Doi: 10.1093/ecco-jcc/jjac059

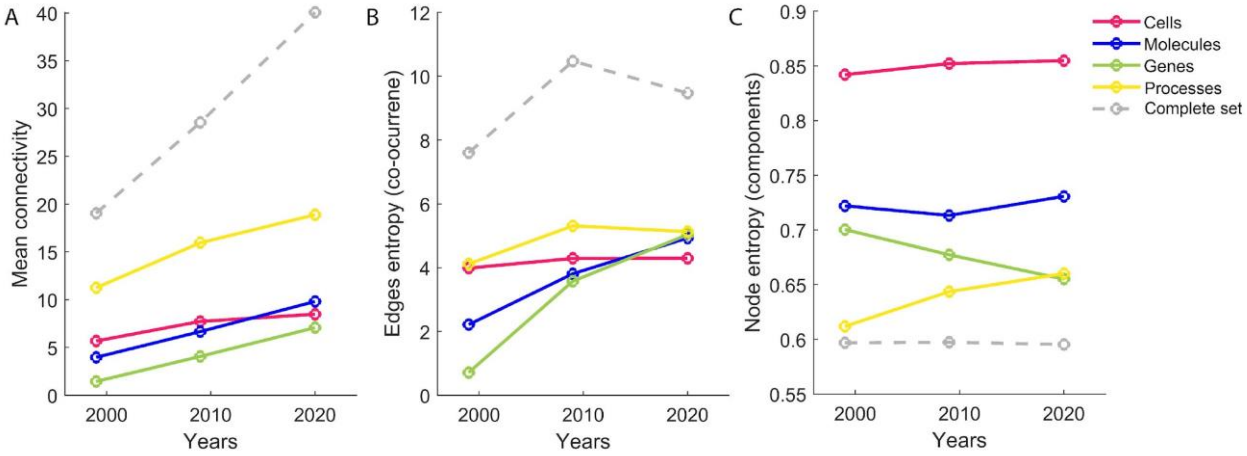



**Fig. 5. Temporal dynamics of the information contained in the articles along decades: comparison among functional levels and the complete set of components.**