*MEDINFO 2023 — The Future Is Accessible*
*J. Bichel-Findlay et al. (Eds.)*

*doi:10.3233/SHTI231122*

# A Text Mining Approach to Explore IFNε Literature and Biological Mechanisms

Mary MCCABE[a], Helen E. GROVES[a], Ultan F. POWER[a] and Guillermo LOPEZ CAMPOS[a,1]
[a] *Wellcome-Wolfson Institute for Experimental Medicine, Belfast, United Kingdom*
ORCiD ID: Mary McCabe https://orcid.org/0000-0002-9307-5399

**Abstract.** Interferons (IFN) constitute a primary line of protection against mucosal infection, with IFN research spanning over 60 years and encompassing a vast ever-expanding amount of literature. Most of what is currently understood has been derived from extensive research defining the roles of "classical" type I IFNs, IFNα and IFNβ. However, little is known regarding responses elicited by less well-characterized IFN subtypes such as IFNε. In this paper, we combined a deductive text mining analysis of IFNε literature characterizing literature-derived knowledge with a comparative analysis of other type I and type III IFNs. Utilizing these approaches, three clusters of terms were extracted from the literature covering different aspects of IFNε research and a set of 47 genes uniquely cited in the context of IFNε. The use of these "in silico" approaches support the expansion of current understanding and the creation of new knowledge surrounding IFNε.

**Keywords.** Text-mining, bioinformatics, interferons, interferon epsilon

## 1. Introduction

Interferons (IFNs) are crucial innate immune mediators against viral infection, produced by virally infected and/or stimulated cells to elicit dynamic antiviral responses. Primarily, type I (IFN-α, -β, -ε, -κ, -τ, -δ, - ζ and -ω) and III IFNs (IFNλ1, -λ2, -λ3 and λ4) are essential in restricting viral replication and dissemination [1]. Our current understanding of IFN expression and signaling is derived from an assessment of the literature surrounding the well-characterized *in vitro* and *in vivo* antiviral activities of IFN-α, -β and -λ [2]. In contrast, there are major literature gaps regarding other IFN subtype expression, induction, signaling and effector capabilities. For example, interferon epsilon (IFNε) is a highly conserved type I IFN, with literature describing tissue and cell line-specific anti-microbial activities *in vitro* and *in vivo* [3,4]. However, unlike other "classical" type I IFNs, induced anti-microbial responses are less potent and its expression appears to be independent of infection and/or pathogen recognition receptor stimulation [5]. Instead, IFNε is a "non-classical" type I IFN, expressed constitutively at multiple mucosal sites, possibly regulated by distinct mechanisms in a cell/tissue-specific manner. Therefore, as a result of its overlapping yet distinctive properties, our understanding of the functions of IFNε is limited.

In the last couple of decades, advances in natural language processing techniques (NLP), together with the vast amount of data and information available in textual format

[1] Corresponding Author: Guillermo Lopez Campos, Address: Wellcome-Wolfson Institute for Experimental Medicine, 97 Lisburn Road, Belfast, BT9 7BL, United Kingdom, email: g.lopezcampos@qub.ac.uk.

have enabled text mining analysis of bibliographic resources to retrieve and investigate data and information available in scientific literature. An example of these applications is LitCovid which extracts and presents updated information in the complex and rich scenario of SARS-CoV-2 research [6]. Another area where text mining applications are used is in the characterization of relatively novel or under-explored research topics where the automatic extraction of multiple biomedical features from the text offers in-depth information about different aspects such as genes, phenotypes, or diseases.

In this study, we present the results of a comparative "*in silico*" analysis of the literature related to IFNε, particularly focusing on information retrieval and its utilization to infer biological mechanisms associated with IFNε, identification of potential diseases where IFNε might play a role and a comparison with other type I and type III IFNs.

## 2. Methods

The methodology used in this paper covers the use of two different text mining-based approaches. The first focused on information retrieval and the characterization of the literature around IFNε. The second combined information retrieval and bioinformatics analyses for the comparison of the different types of IFNs. Detailed descriptions of each of the approaches are provided below.

For the first analysis, we built the following PubMed query "Interferon epsilon" or "Interferon Epsilon" or "IFNE" or "IFNε" or "IFN-E" or "IFN-ε" or "Interferon-ε" or "IFN-epsilon"" to retrieve available abstracts. Analyses were done in R (v4.2.1) using packages "RISmed", "tm", "SnowballC" and "wordcloud" [7-10]. Briefly, PMID values were given to "RISMed" to import publications into R, creating a research corpus where they were then further preprocessed using "tm", "SnowballC" and finally, "wordcloud" for data visualization. A complementary analysis focused on the identification of terms, concepts built from two or more words, was completed using TerMine. This facilitated the identification of high-frequency multi-word terms within the IFNε abstract corpus ranked based on length, frequency and longer-term incorporation (C-score) [11]. Finally, to complete this descriptive work around IFNε literature, VOSviewer (version 1.6.19) was used to build a co-occurrence network within identified publications using total link strength (TLS) [12] and identify clusters that can be associated with different topics identified from the literature.

The second approach was based on the use of an in-house developed pipeline for the automatic analyses of the literature available in PubMed [13]. Using Pubtator Central annotations at the core of these methods, four different queries were used based on the first approach query but replacing IFNε with each of the other well-characterized IFNs, IFNα, -β, and -λ. These analyses retrieved information about genes, diseases, and genetic variants from June 1980 to February 2023. A second stage involved the use of the retrieved genes to infer biological mechanisms (pathways) utilizing the functional profiling tool g:Profiler [14].

## 3. Results

A total of 89 articles were returned on the topic of IFNε. The initial word frequency analysis results are displayed in the wordcloud in Figure 1A. TerMine identified 1640 high-frequency multi-word terms mined from the IFNε abstract corpus. Ranked first was

"ifne gene" with a C-value score of 27. Extracted terms also included many associated with the antiviral response and viral infections, such as "antiviral activity" (C-score 23), "sars-cov-2 infection" (C-score 3) and "influenza virus" (C-score 5); terms associated with cancer, such as "cancer risk pathway enrichment analysis" (C-score 2) and "normal mismatch repair" (C-score 2); terms associated with female reproductive organs, such as "female reproductive tract" (C-score 10); and terms associated with the airway mucosae, such as "nasal epithelial cell" (C-score 2) and "olfactory signaling pathway" (C-score 2). VOSviewer identified a total of 2886 terms, of which 85 occurred at least 5 times in the abstract corpus. Co-occurring terms grouped into three clusters (Figure 1B). The most frequent co-occurring term was "ifne" (TLS 121) with "viral infection" possessing a TLS 74. The term "ifne" was also linked to "epithelial cell" (TLS 53), "lung" (TLS 38) and "mucosal immunity" (TLS 37).

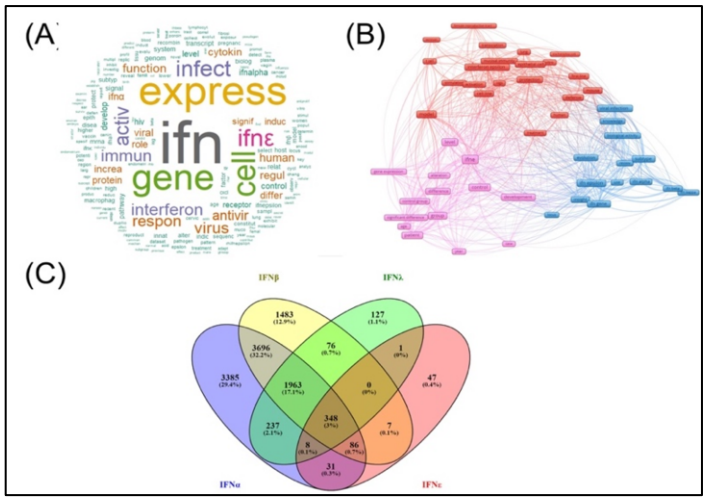


**Figure 1.** Short caption Most frequent term, term co-occurrence, and text mining analysis of IFNε-only and IFNα, -β, -λ and -ε publications. (A) Wordcloud displaying the most frequent terms extracted from IFNε-only abstracts (B) Network visualization of abstract derived term co-occurrence with the 3 clusters C) Venn diagram presenting the complex landscape of genes identified in the comparative analyses of IFNα, -β, -λ and -ε. Full size/Hi-res images are available at 10.6084/m9.figshare.22193614.

The comparative analyses of the four types of different IFNs showed that IFNα could be annotated with a larger number of uniquely cited genes (3385), followed by IFNβ (1483) and IFNλ (127) (Figure 1C). Of the total 528 IFNε annotated genes, 348 were shared with IFNα, -β and -λ, including type I, II and III IFNs and classical inflammatory cytokine and interferon stimulated genes, such as CXCL8, IL1β, ISG15 and MX1. Forty-seven genes were unique to IFNε-queried abstracts, including a nasal epithelial marker (BPIFA1), a folate receptor associated with assorted epithelial malignancies (FOLR3), and an olfactory receptor (OR2L8).

Flat gene lists acquired from text mining each IFN query were submitted to g:Profiler to facilitate pathway inference (using a significance threshold of Benjamini-Hochberg adj.p. Value <0.05). Amongst those shared between all IFNs, archetypal IFN and antiviral pathways, such as "JAK-STAT signaling pathway" (KEGG:04630) and "SARS-CoV infections" (REAC: R-HSA-9679506) were present. Inferred pathways unique to IFNε gene annotations included "Antimicrobial peptides" (REAC: R-HSA-168249), "neoplasia of the nasopharynx" (HP:0001739), and "Diseases of Mismatch

Repair (MMR)" (REAC: R-HSA-9675135). A subset of pathways referring to lung development, such as "mesenchymal cell proliferation involved in lung development" (GO:0060916) and "lung goblet cell differentiation" (GO:0060480), were also apparent.

## 4. Discussion

We devised a strategy to study IFNε using biomedical literature and a two-stage biomedical informatics methodology. An initial deductive approach was able to assess the current landscape of IFNε literature, combining text analysis and mining approaches to extract IFNε-specific knowledge. The high level of co-citation and cross-talk across the different IFNs necessitated the use of abstracts rather than full text for our analyses to reduce the detection of potential false-positives.

Although the initial characterization of the 89 IFNε-associated publications identified in this study would have allowed for a manual literature assessment, this would have been hampered by the need of highly trained annotators and would have focused only on the extraction of certain aspects (such as genes) from the literature. In addition, for the assessment of IFNε in the context of other type I and type III IFNs a computational approach was required as it would have implied the analysis of more than 40 years of research and tens of thousands of documents. Most frequent term extraction and co-occurrence analysis highlighted that IFNε was most frequently discussed in the context of its antiviral function and with other type I IFN subtypes, its role in the female reproductive tract (FRT) and, interestingly, its less well-characterized non-FRT epithelial cell and cancer associations. VOSviewer analyses identified three different clusters (Figure 1B) associated with the different biological aspects covered in the IFNε literature, with clusters highlighting the context of its type I IFN status (blue), gene expression analyses (pink) and its biological role (red). More complex text mining approaches, such as topic modelling, could be used in the future, benefiting from additional research, to provide a better unsupervised approach and identification of other areas of research.

The in-house analytical pipeline querying and comparing the multiple IFN-associated abstracts enabled the identification of distinct INFε related genes. Interestingly, these included genes related to areas to be further explored in the context of INFε, such as its possible airway mucosal expression, in parallel with genes linked with varied tissue-type epithelial cell cancers. However, as expected as a type I IFN, the greatest overlap of IFNε-associated gene annotations was with IFNα, -β and -λ annotations, which included many antiviral and innate immune signaling pathway components, emphasizing the central antiviral role of IFNε within the current literature. A significant step in our approach is the development of new knowledge and hypothesis generation derived from pathway inference analyses. This enabled the identification of INFε unique inferred pathways. Moreover, our approach allows the comparison of pathways inferred from text-mined gene annotations to those identified via TerMine providing supporting evidence aiding in the validation of this study's multi-query text-mining approach. The combination of these two different but complementary approaches can back each other's findings and allow the identification of novel pathways biologically plausible in the context of IFN and INFε and creating novel hypotheses to be validated experimentally.

## 5. Conclusions

The use of biomedical informatics tools for the study of IFNε and its comparison with other interferons has allowed a detailed description of the knowledge around this area. The use of biomedical informatics tools to extract information and generate new knowledge from pathway inference allows us not only to improve our understanding of the role and mechanisms associated with this under-studied interferon but also paves the way for the development of novel knowledge bases around these important cytokines. IFNε is associated with and shares many attributes with classical type I and III IFNs. However, in this study we utilized knowledge, text mining and pathway inference approaches to facilitate the disentanglement of current interferon knowledge, identifying IFNε-specific associated attributes and aiding in the exploration of the "non-classical" features of this poorly understood interferon.

## References

[1] Müller U, Steinhoff U, Reis LF, Hemmi S, Pavlovic J, Zinkernagel RM, Aguet M. Functional role of type I and type II interferons in antiviral defense. Science. 1994 Jun;264(5167):1918-21, doi: 10.1126/science.8009221.
[2] Lazear HM, Schoggins JW, Diamond MS. Shared and distinct functions of type i and type iii interferons. Immunity. 2019 Apr;50(4):907-23, doi: 10.1016/j.immuni.2019.03.025.
[3] Mungin JW, Chen X, Liu B. Interferon epsilon signaling confers attenuated zika replication in human vaginal epithelial cells. Pathogens. 2022 Jul;11(8):853, doi: 10.3390/pathogens11080853.
[4] Bourke NM, Achilles SL, Huang SU, Cumming HE, Lim SS, Papageorgiou I, Gearing LJ, Chapman R, Thakore S, Mangan NE, Mesiano S, Hertzog PJ. Spatiotemporal regulation of human IFN-ε and innate immunity in the female reproductive tract. JCI Insight. 2022 Sep;7(18):e135407, doi: 10.1172/jci.insight.135407.
[5] Fung KY, Mangan NE, Cumming H, Horvat JC, Mayall JR, Stifter SA, De Weerd N, Roisman LC, Rossjohn J, Robertson SA, Schjenken JE, Parker B, Gargett CE, Nguyen HP, Carr DJ, Hansbro PM, Hertzog PJ. Interferon-ε protects the female reproductive tract from viral and bacterial infection. Science. 2013 Mar;339(6123):1088-92, doi: 10.1126/science.1233321.
[6] Chen Q, Allot A, Lu Z. LitCovid: an open database of COVID-19 literature. Nucleic Acids Res. 2021 Jan;49 (D1):D1534–40, doi: 10.1093/nar/gkaa952.
[7] Kovalchik S. RISmed 2016. Available from: https://CRAN.R-project.org/package=RISmed.
[8] Feinerer I, Hornik K, Meyer D. Text mining infrastructure in R. J Stat Softw. 2008 Mar;25(5):1-54, doi: 10.18637/jss.v025.i05
[9] Bouchet-Valat M. SnowballC: Snowball stemmers based on the Clibstemmer UTF-8 library. 2014. Available from: https://cran.r-project.org/package=SnowballC
[10] Fellows I. wordcloud: Word Clouds. 2014. Available from: https://cran.r-project.org/package=wordcloud.
[11] Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms:.the C-value/NC-value method. Int J Digit Libr. 2000 Aug;3:115-13, doi: 10.1007/s007999900023.
[12] van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics. 2010 Aug;84(2):523-38, doi: 10.1007/s11192-009-0146-3.
[13] Lopez-Campos G, Bonner E, McClements L. An integrative biomedical informatics approach to elucidate the similarities between pre-eclampsia and hypertension. Stud Health Technol Inform. 2019 Aug;264:988-92, doi: 10.3233/SHTI190372.
[14] Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019 Jul;47(W1):W191-8, doi: 10.1093/nar/gkz369.