Extracting T Cell Function and Differentiation Characteristics from the Biomedical Literature

Eric Czech[1§], Jeff Hammerbacher[1]

[1] Microbiology and Immunology Department at Medical University of South Carolina
[§] Corresponding author

Abstract

The role of many cytokines and transcription factors in the function and development of human T cells has been the subject of extensive research, however much of this work only demonstrates experimental findings for a relatively small portion of the molecular signaling network that enables the plasticity inherent to these cells. We apply recent advancements in methods for weak supervision of natural language models to aid in extracting these individual findings, as ~400k cell type, cytokine, and transcription factor relations from ~10k full-text PMC articles, before collating them as freely available datasets, analysis and source code available at https://github.com/hammerlab/t-cell-relation-extraction.

Introduction

Many promising cancer immunotherapy treatment protocols rely on efficient and increasingly sophisticated methods for manipulating human immune cells. T cells are a frequent target of the laboratory and clinical research driving the development of such protocols as they are most often the effector of the cytotoxic activity that makes these treatments so potent. However, the cytokine signaling network that drives the differentiation and function of such cells is complex and difficult to replicate on a large scale in model biological systems. Abridged versions of these networks have been established over decades of research but it remains challenging to define their global structure as the classification of T cell subtypes operating in these networks, the mechanics of their formation, and the purpose of the signaling molecules they excrete are all controversial, with a slowly expanding understanding emerging in literature over time.

We present an automated method for extracting links in these networks and then apply this method to a large corpus of PMC articles relating to immunology. All of the links extracted relate exclusively to T cell subtypes and include cytokines that either induce differentiation (hereinafter referred to as “inducing cytokines”) of these subtypes or are secreted by them (“secreted cytokines”), as well as the transcription factors involved in the path along any developmental trajectory (“inducing transcription factors”). This former type of link, between T cell subtypes and secreted or inducing cytokines, was quantified over a large corpus and shared first in the

immuneXpresso [1] database, which also includes a much larger breadth of cell types and covered the entirety of PubMed abstracts available at the time. A primary goal in this study is to be able to recapitulate much of this data while utilizing more expressive language models trained with less manual annotation effort, a capability conferred by recent advancements in weak supervision [2]. We also extend these efforts to incorporate full text articles, which provide access to far more relations than PubMed abstracts, as well as transcription factors (novel contributions) after lending veracity to the findings through comparison with the shared findings first published in immuneXpresso. In summary, this study is intended to demonstrate the viability of weak supervision for biological relation extraction in scientific literature as well as share a large database of T cell-specific cytokine and transcription factor relationships.

Results

We trained and extracted 3 different relations with examples of each shown in **Figure 1**. A single LSTM classifier was trained for each relation and applied to a larger corpus as a means of producing a large number of *(cell type, cytokine)* or *(cell type, transcription factor)* tuples with a probability used as threshold for favoring prediction precision over recall. Example high probability predictions can be seen in **Supplementary Figure 8** and word clouds associated with these same predictions can be seen in **Supplementary Figure 9**. These visualizations provide some insight into what text patterns or words the models favor most highly in the identification of representative candidates.

Relation Type	Examples
a) Inducing Cytokine	1. **IL-6** is one of the first cytokine signals to influence the early stages of **Tfh** differentiation 2. **TGF-β** is required for formation of **Th17** cells 3. Th2 development may be due to the suppression of Th1 because of **IL-12** mediated **Th1** differentiation
b) Secreted Cytokine	1. AhR increases the proliferation of **TH17** cells that produce **IL-17** 2. The blockade of Notch1 resulted in the downregulation of **Th17** cells, effective cytokine **IL-17** secretion, and Th17-mediated disease progression 3. **Th2** cells produce **interleukin-4**
c) Inducing Transcription Factor	1. These compounds were effective in blocking **RORγt** driven luciferase expression and inhibiting **Th17** differentiation but not proliferation 2. The **Th1** subset is defined by expression of the lineage-determining transcription factor **T-bet** 3. **forkhead box p3**, a signature transcription factor of **regulatory T** cells
Figure 1: Relation example instances with an individual candidate relation highlighted as either a cytokine or transcription factor in red and a T cell subset name in blue. These	

examples show a few of the difficulties that arise in classifying links made through indirect language (such as in **c1**) or noun-phrases (such as “IL-12 mediated” in **a1**).

Assuming that a probability threshold of 50% is indicative of a valid relation, we summarized all passing tuples to help demonstrate the per-cell type cytokine profiles that can be inferred based on the frequency of each tuple in **Figure 2**. These profiles agree well with existing literature [3] on the cytokines most commonly associated with the function and differentiation of many T cell subtypes. Unique document counts associated with each relation are shown in **Figure 3** which also provides are better sense of the cell types involved in this study as well as the breadth of cytokines and transcription factors encountered.

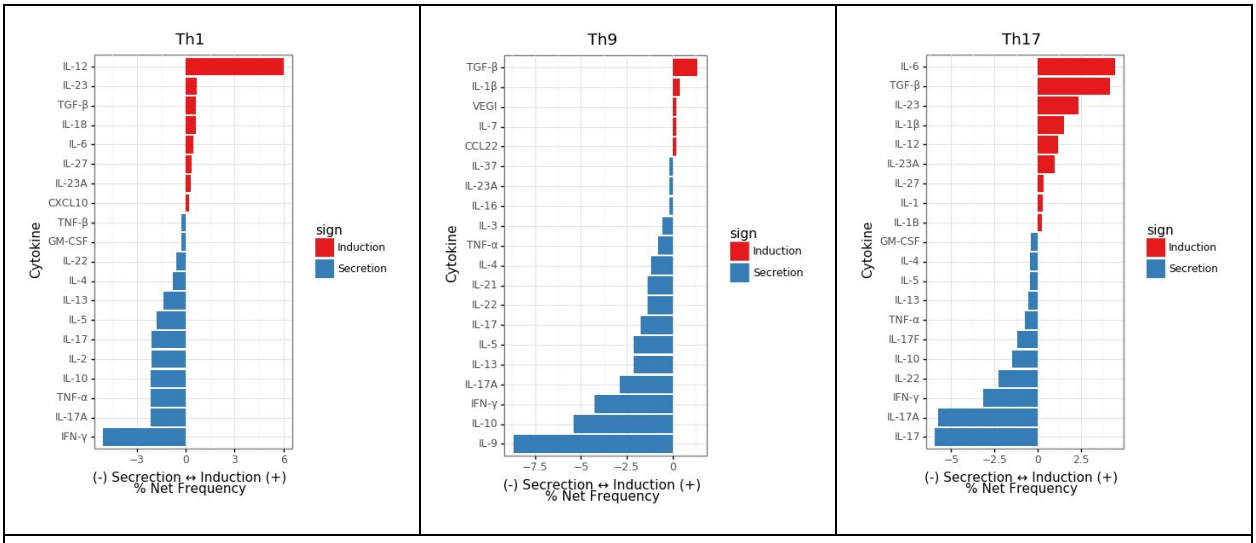
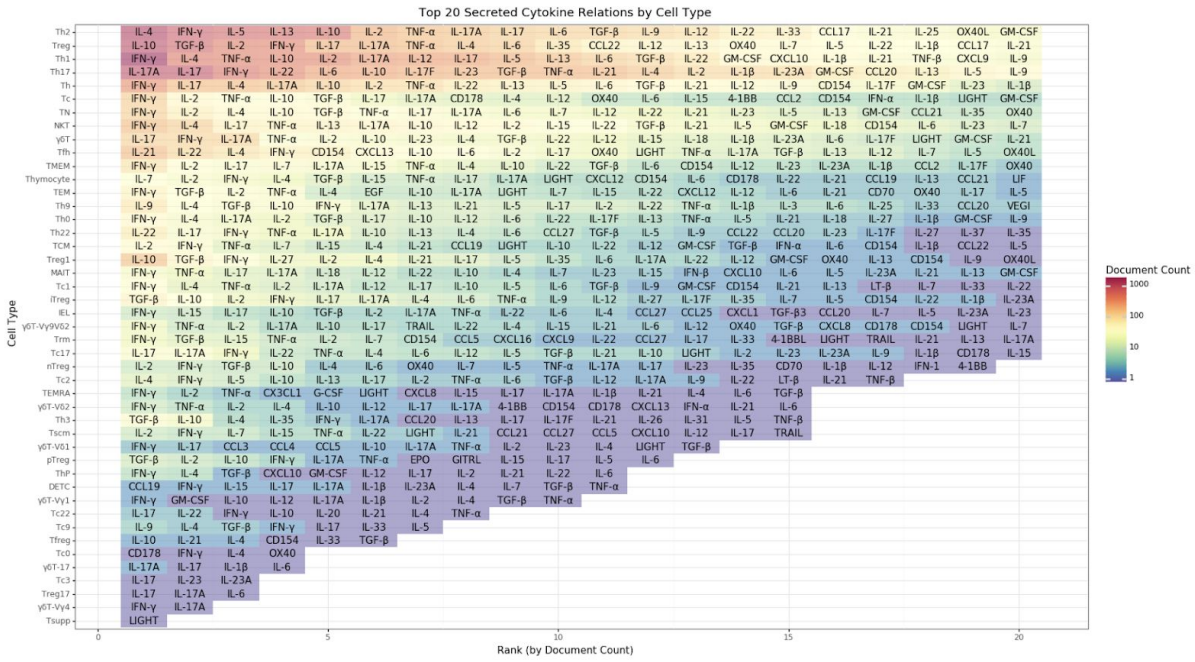
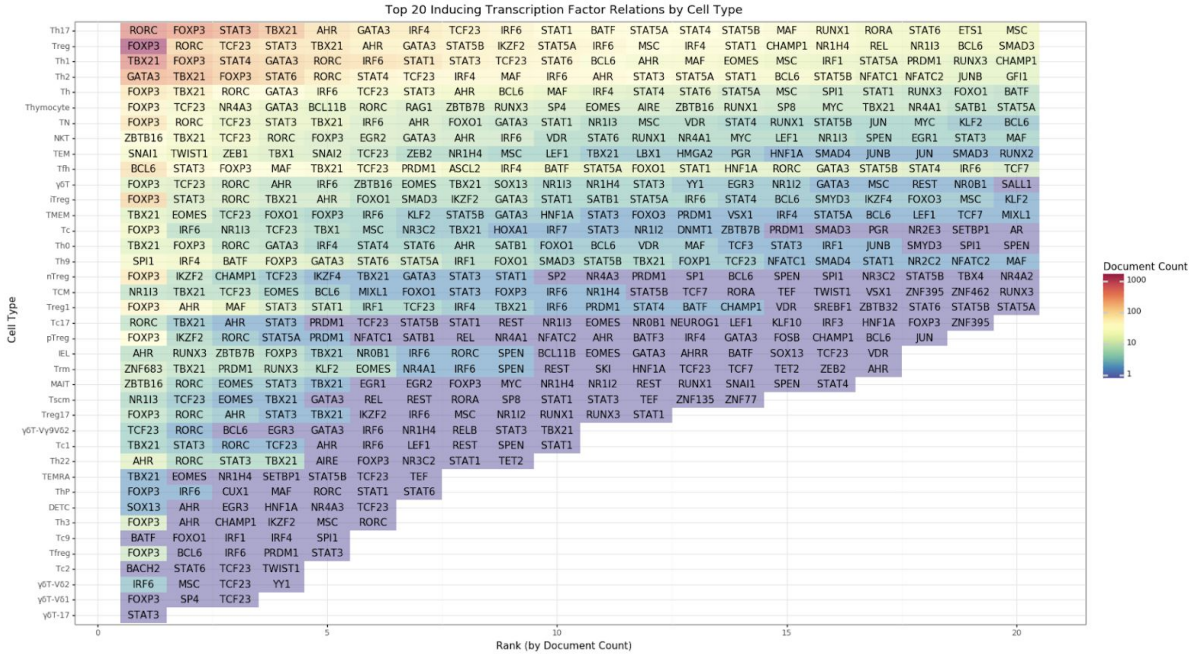


Figure 2: Difference in frequencies of secretion and induction relations for most relevant cytokines to each of three different T cell types. Cytokines with a positive sign (red) most frequently occur within the context of cell differentiation while those with a negative sign (blue) are most often referenced in regards to the functional profile of the cell type.


Top 20 Secreted Cytokine Relations by Cell Type
Cell Type
Rank (by Document Count)
Document Count
1000
100
10
1


Top 20 Inducing Transcription Factor Relations by Cell Type
Cell Type
Rank (by Document Count)
Document Count
1000
100
10
1

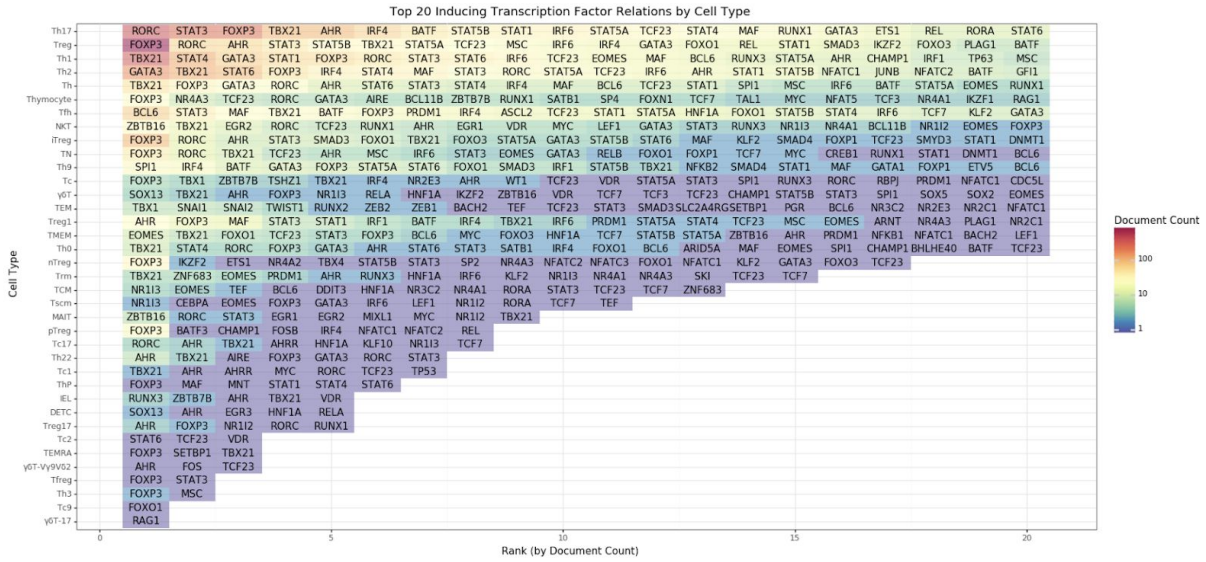
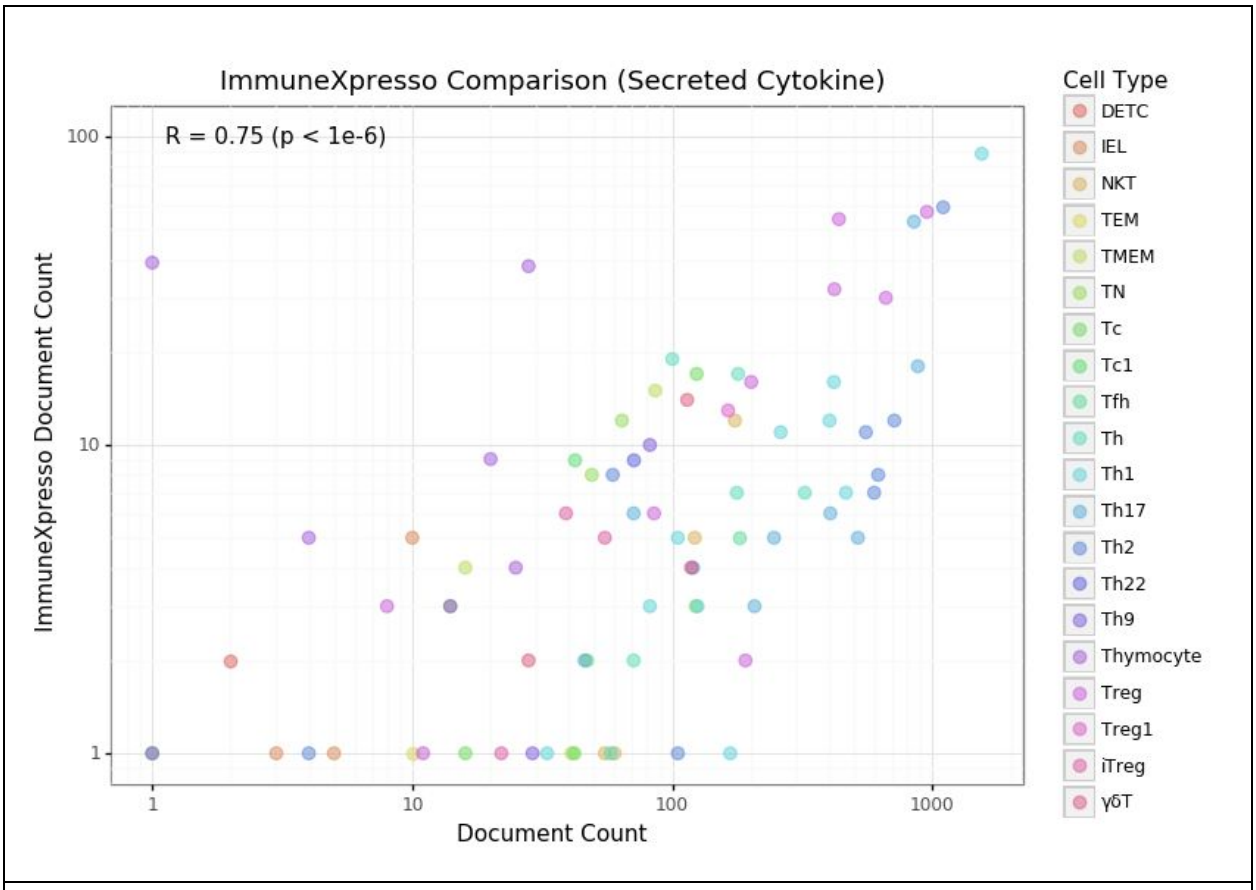


Figure 3: Top cytokines and transcription factors for each T cell type as determined by the number of unique documents in which a high probability (p > .5) relation instance occurs.

For further validation of T cell subtype and cytokine relations, we also compared our extractions to those of [1], which are summarized in **Figure 4** as well as **Figure 5**. **Figure 4** shows an important distinction in our methodology in that the document count for each cell type and cytokine pair is far higher, due to the use of full text documents rather than just abstracts. **Figure 5** also shows alignment in the methodologies by demonstrating that conducting a similar analysis on secreted/inducing cytokine links leads to very similar results.


ImmuneXpresso Comparison (Secreted Cytokine)
R = 0.75 (p < 1e-6)
ImmuneXpresso Document Count
Document Count
100
10
1
1
10
100
1000
Cell Type
DETC
IEL
NKT
TEM
TMEM
TN
Tc
Tc1
Tfh
Th
Th1
Th17
Th2
Th22
Th9
Thymocyte
Treg
Treg1
iTreg
γδT

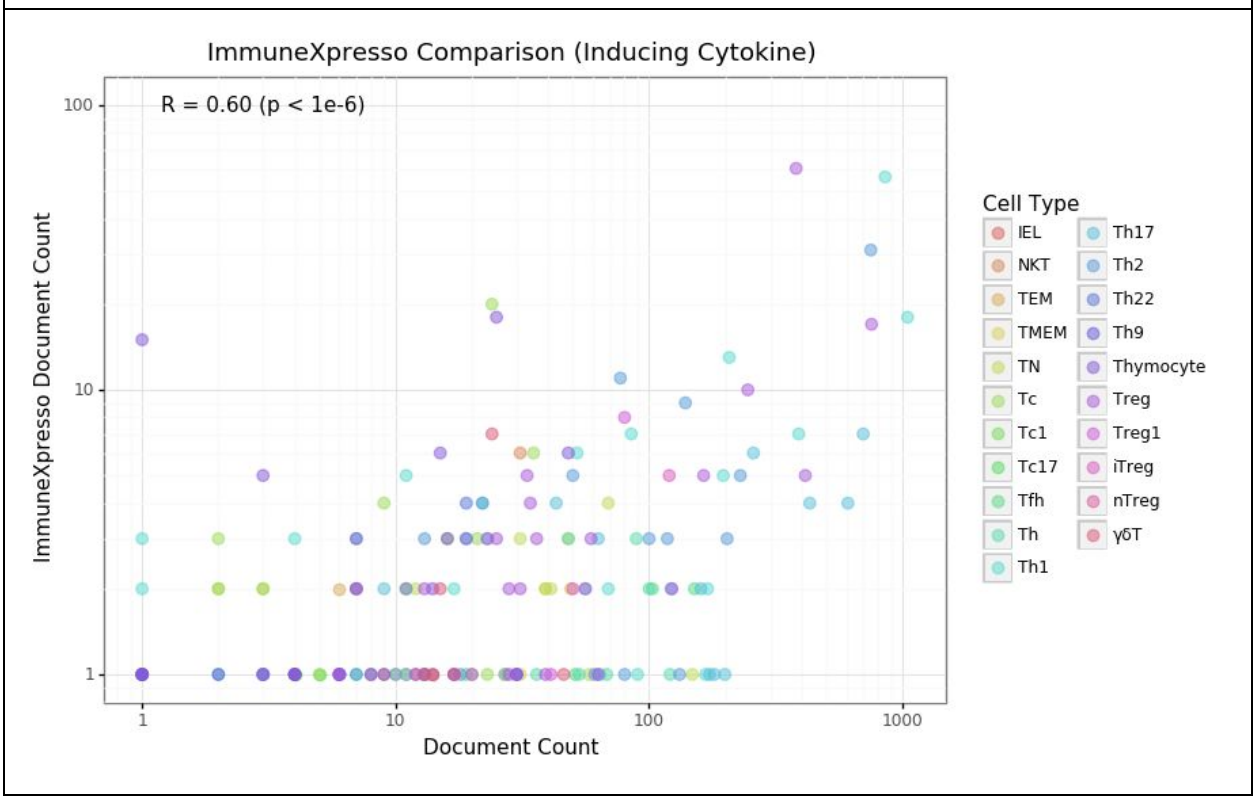

ImmuneXpresso Comparison (Inducing Cytokine)
R = 0.60 (p < 1e-6)
ImmuneXpresso Document Count
Document Count
100
10
1
1
10
100
1000
Cell Type
IEL
NKT
TEM
TMEM
TN
Tc
Tc1
Tc17
Tfh
Th
Th1
Th17
Th2
Th22
Th9
Thymocyte
Treg
Treg1
iTreg
nTreg
γδT

Figure 4: ImmuneXpresso comparison showing the number of documents discovered for each cell type and cytokine relation that was also present within this study (82 such pairs for secreted cytokine relation and 185 for inducing cytokine relation). Pearson correlation in top left shows modest but significant agreement (two-tailed t-test).

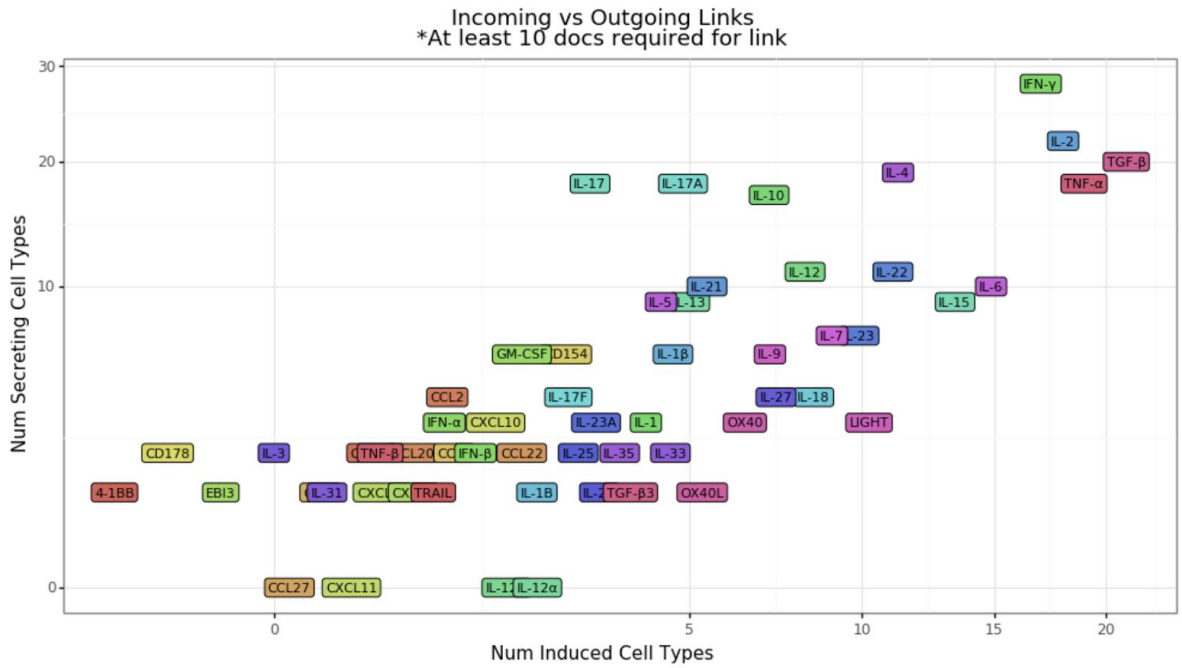


Figure 5: Comparison of links between cytokines and cell types that they either induce or are secreted by (or both). A single link in this case is only included when found in at least 10 distinct documents to ensure that the link is more likely to be valid. This is comparable to Figure 2c of [1] where the Pearson correlation between the number of links was found to be .86, very similar to the .87 found here.

Trends in time were also observable, particularly with respect to the predominance of certain T cell types within the literature targeted in this study. Mentions of these types are shown in **Figure 6** where it is clear that regulatory T cells, and the closely related Th17 cells, have received a large increase in attention over the last 13 years. The discovery and acceptance of now putative cell types such as follicular regulatory T, Th9, Th22, and EMRA (effector memory cells re-expressing CD45RA) T cells is also observable within this time span.

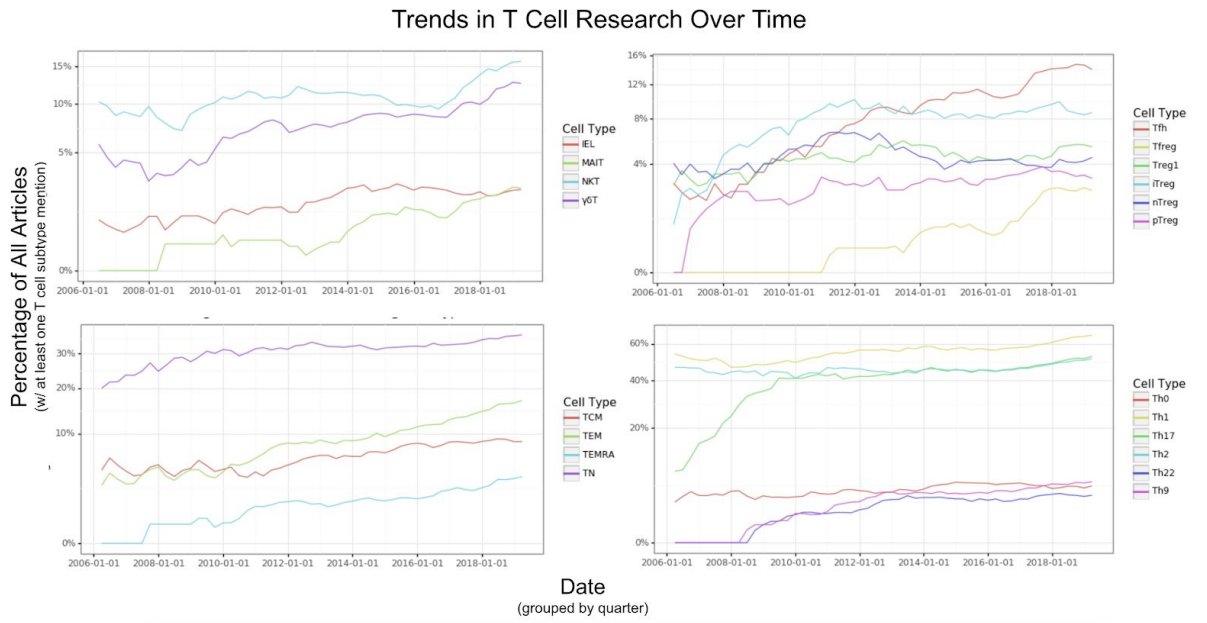


Figure 6: Trends in mentions of particular T cell types over time (previous 13 years) shown as percentages of all 7,638 documents mentioning at least one T cell type included in this study. All values are shown as rolling averages over an 8 quarter (2 year) period.

Methods

Document Collection

The collection of documents for this study involved batched Entrez queries for PMC articles matching a boolean keyword query[1]. This query matched 120k documents, however, only 10k were included in this study with approximately 70% of those documents having full-text content available (~7k of 10k). Text for processing was formed as the concatenation of each document's title, abstract, and full-text content when possible.

Tagging

Entity tagging for cytokines, transcription factors, and cell types was applied using known symbol and alias lookup tables[2]. Cytokine and Transcription Factor names as well as associated aliases were integrated primarily from [4] with a large number of transcription factor aliases also

[1] *PMC article query: (human) AND ((t cell) OR (t lymphocyte)) AND (cytokine) AND ((differentiate) OR (differentiation) OR (differentiated)) AND ((polarization) OR (polarize) OR (induce) OR (induction)*

[2] These tables are available at https://github.com/hammerlab/t-cell-relation-extraction/tree/master/pm_subtype_protein_relations/data/meta

provided by [5]. Cell types and aliases were gathered as curated lists from applications of the JNLPBA [6] named entity recognition tagger provided in SciSpacy [7].

Training

Weak supervision for the relation classifiers was conducted as outlined in **Figure 7** where several different strategies were used to construct labeling functions. The first of these strategies included noting text patterns observed in annotation of a labeled relation set for performance evaluation. Examples of these patterns[3] include generalized expressions of the nature of the relationship between a particular cell type and protein (e.g. “{cytokine} is (essential|important|critical) for {cell type} (development|differentiation)”). Common expressions indicative of an inhibitory relationship between a certain protein and cell type were also matched against (e.g. “{transcription factor} is critical for (suppressing|inhibiting|repressing) (induction|differentiation|development) of {cell type} cells)”). Providing many such patterns along with a single source of distant supervision, ImmuneXpresso, and a few structural heuristics allowed for the classifiers to be fit without any manually annotated training data. This training data included 11.7k inducing/secreted cytokine relation candidates, as the candidates for these two relations are the same and the relation type could be applicable simultaneously, as well as 6.7k inducing transcription factor relation candidates. All models were defined as 100 node LSTM networks with an equally sized embedding dimensionality (100) and were trained for 20 epochs with a 50% dropout rate and rebalancing to give a 25% to 75% ratio in positive to negative outcomes.

[3] All patterns applied can be found at https://github.com/hammerlab/t-cell-relation-extraction/blob/snorkel_relations/pm_subtype_protein_relations/src/supervision.py

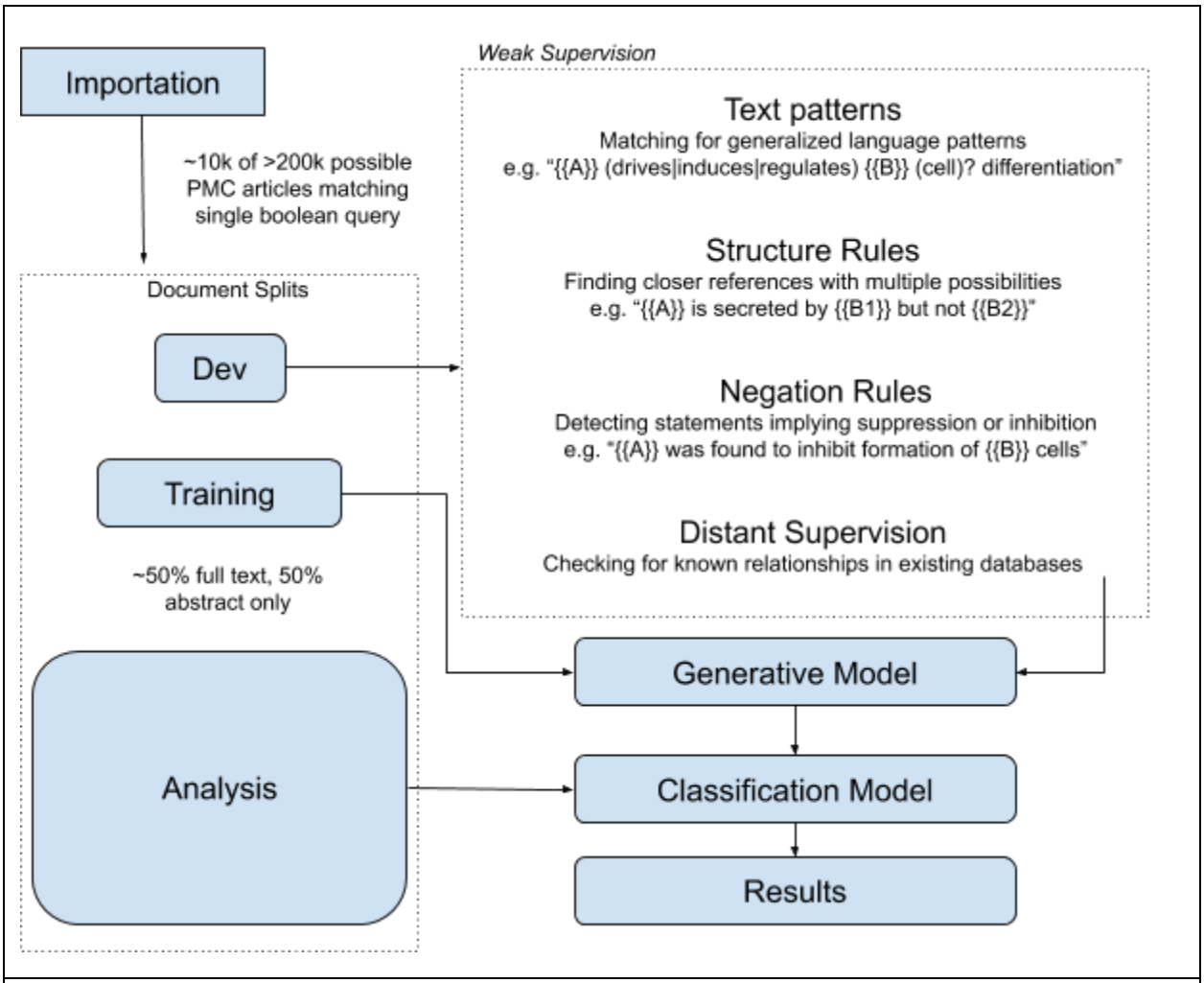


Figure 7: Data collection and training process summary.

Conclusion

Weak supervision enabled extraction of biomedical entity relationships in this study with much less annotation effort than that seen in the previous work by [1]. Results were also found to be comparable for shared relation types in the two studies. Additionally, the inclusion of additional relationships between T cell types and transcription factors in this study may serve as a useful collation of likely links in polarization pathways, crucial information for the construction of experiments in transfection-based cell fate manipulation.

Supplementary Material

Inducing Cytokine:

77.39%| **IL-6** induces the differentiation of **Th17** cells from naïve T cells through STAT3

76.59%| Nevertheless, the involvement of p38 MAPK in **GITRL** induced **Th17** cell differentiation has not been described

75.40%| Another study in mice further indicated that **IL-23** was required for driving terminal **Th17** differentiation.

Secreted Cytokine:

75.12%| **IL-17** is mainly produced by **Th17** cells, macrophages, and neutrophils.

74.87%| Importantly, Blimp-1-dependent **IL-10** production by **Tr1** cells is a major regulator of tumor necrosis factor (TNF)-mediated inflammation

74.51%| **Th17** cells express a transcription factor ROR-γt and produce cytokines **IL-17** and IL-22

Inducing Transcription Factor:

97.86%| **Sox5** and c-Maf cooperatively induce **Th17** cell differentiation

97.84%| Overexpression of **GATA3** in naive CD4 T cells causes **Th2** differentiation even in the absence of IL-4.

97.84%| **ETV5** axis regulates **TFH** cell differentiation via MAF.

Supplementary Figure 8: Example high probability relation classifications for each relation type.

Supplementary Figure 9: Word clouds for most and least likely 250 sentences corresponding to each relation class.

References

1. Kveler K, Starosvetsky E, Ziv-Kenet A, Kalugny Y, Gorelik Y, Shalev-Malul G, et al. Immune-centric network of cytokines and cells in disease context identified by computational mining of PubMed. Nat Biotechnol. 2018;36: 651–659.

2. Ratner A, De Sa C, Wu S, Selsam D, Ré C. Data Programming: Creating Large Training Sets, Quickly. Adv Neural Inf Process Syst. 2016;29: 3567–3575.

3. Raphael I, Nalawade S, Eagar TN, Forsthuber TG. T cell subsets and their signature cytokines in autoimmune and inflammatory diseases [Internet]. Cytokine. 2015. pp. 5–17. doi:10.1016/j.cyto.2014.09.011

4. Xin J, Mark A, Afrasiabi C, Tsueng G, Juchler M, Gopal N, et al. High-performance web services for querying gene and variant annotation. Genome Biol. 2016;17: 91.

5. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription Factors. Cell. 2018;172: 650–665.

6. Kim J-D, Ohta T, Tsuruoka Y, Tateisi Y, Collier N. Introduction to the bio-entity recognition task at JNLPBA [Internet]. Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications - JNLPBA ’04. 2004. doi:10.3115/1567594.1567610

7. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing [Internet]. 2019. Available: http://arxiv.org/abs/1902.07669