

Immune-centric network of cytokines and cells in disease context identified by computational mining of PubMed

Ksenya Kveler¹, Elina Starosvetsky¹, Amit Ziv-Kenett^{1,7}, Yuval Kalugny^{1,2}, Yuri Gorelik¹, Gali Shalev-Malul¹, Netta Aizenbud-Reshef¹, Tania Dubovik¹, Mayan Briller¹, John Campbell³, Jan C Rieckmann^{4,7}, Nuaman Asbeh¹, Doron Rimar^{1,5}, Felix Meissner⁴, Jeff Wiser^{3,7} & Shai S Shen-Orr^{1,6}

Cytokines are signaling molecules secreted and sensed by immune and other cell types, enabling dynamic intercellular communication. Although a vast amount of data on these interactions exists, this information is not compiled, integrated or easily searchable. Here we report immuneXpresso, a text-mining engine that structures and standardizes knowledge of immune intercellular communication. We applied immuneXpresso to PubMed to identify relationships between 340 cell types and 140 cytokines across thousands of diseases. The method is able to distinguish between incoming and outgoing interactions, and it includes the effect of the interaction and the cellular function involved. These factors are assigned a confidence score and linked to the disease. By leveraging the breadth of this network, we predicted and experimentally verified previously unappreciated cell–cytokine interactions. We also built a global immune-centric view of diseases and used it to predict cytokine–disease associations. This standardized knowledgebase (<http://www.immunexpresso.org>) opens up new directions for interpretation of immune data and model-driven systems immunology.

Protective immunity is mediated through a complex system of interacting cells whose communication network is primarily governed by secreted molecules, chiefly members of the cytokine and chemokine family of proteins. Until recently, the high complexity of the immune system was approached by researchers by using reductionist approaches, but technological advances now enable acquisition of large data sets, with broad enumeration of cell subset types and functions, proteins, gene expression and more¹. In addition, papers in immunology alone are being published at the rate of approximately one every 30 min. To maximize discovery, research results must transition to organized standardized models of knowledge, on which automated computational processing can be deployed.

Biomedical text-mining efforts have been an important means of grasping at the breadth and complexity of biological systems. With efforts invested into recognizing biologically relevant entities—such as genes, diseases, chemicals and genomic variants^{2–8}—driven by gold standards^{9,10} and community-wide efforts^{11–14}, text-mining is enabling automatic identification of complex biological relations^{15,16} and full-scale networks. Recent research has expanded to additional types of molecular events^{17–19}, with relation-extraction methods ranging from co-occurrence^{15,19}, pattern matching and rule-based methods to dependency parse graph analysis^{20,21} and machine-learning²¹.

To date, however, text-mining approaches have not addressed large-scale intercellular communication networks and, in particular, those describing directional cell–cytokine interactions.

Biological literature mining has shown utility for hypothesis generation, particularly in disease contexts^{22–24}. Similarly, data-driven disease classifications have shown benefit in understanding shared mechanisms, empowering target identification and drug repositioning choices^{25–28}. Yet to date, such classifications have not addressed cellular cross-talk and how the immune system may impact disease.

To establish a foundation for systematic reasoning over the intercellular network, we built immuneXpresso (iX), a comprehensive high-resolution knowledgebase of directional intercellular interactions that was text-mined from all available PubMed abstracts across a broad range of disease conditions. Interactions captured by iX include both direct cytokine binding or secretion events and more distant, indirect influencing relations, scored and filtered to emphasize precision. We used the resulting knowledge standardization to characterize the immune intercellular network and to predict and experimentally validate cell–cytokine interactions. By leveraging the breadth and context-awareness of the knowledgebase, we built an immune-centric view of diseases and explored its modularity to predict cytokine–disease associations.

¹Faculty of Medicine, Technion–Israel Institute of Technology, Haifa, Israel. ²CytoReason, Tel-Aviv, Israel. ³Northrop Grumman IT Health Solutions, Rockville, Maryland, USA. ⁴Experimental Systems Immunology, Max Planck Institute of Biochemistry, Bayern, Germany. ⁵Rheumatology Unit, Bnai Zion Medical Center, Haifa, Israel. ⁶Faculty of Biology, Technion–Israel Institute of Technology, Haifa, Israel. ⁷Present addresses: Augury Systems Ltd., Haifa, Israel (A.Z.-K.); Roche Diagnostics GmbH, Penzberg, Germany (J.C.R.); Medidata Solutions, New York, New York, USA (J.W.). Correspondence should be addressed to S.S.S.-O. (shenorr@technion.ac.il).

Received 16 May 2017; accepted 5 April 2018; published online 18 June 2018; doi:10.1038/nbt.4152

RESULTS

A text-mining pipeline to extract intercellular interactions

We designed a computational pipeline that was focused on mining the primary literature for identification of cells, intercellular signaling molecules (i.e., cytokines) and the directional relations between them (**Fig. 1a** and Online Methods), and we applied it across the entire PubMed (approximately 16 million articles published electronically by July 2017). Briefly, for each individual sentence, the analysis pipeline tagged cells, cytokines and diseases, as well as standardized terminology through official ontologies to allow for hierarchical data analysis at multiple resolutions (**Supplementary Tables 1–4**). We examined sentence structure to identify syntactically related cell, cytokine and verb mentions. For each such ‘evidence record’, we detected the relation’s directionality, polarity (representing its positive, negative or neutral effect) and when possible, the resulting cellular biological function (**Supplementary Table 5**). We distinguished between ‘outgoing’ relations (describing cytokine secretion by a given cell type) and ‘incoming’ relations (describing events in which a cytokine affected a cell type, either directly via binding or indirectly). Finally, for each unique triple of cell, cytokine and directionality summarized across all of the evidence records, we used a trained machine-learning classifier to make a call on whether the collected evidence indeed described an interaction (Online Methods). We assigned confidence scores to these and linked them to the conditions (e.g., diseases) co-mentioned in the same abstracts. In addition, we annotated independent entity mentions, without interaction, of cells and cytokines to allow for entity co-occurrence and enrichment statistics.

To assess the precision of entity recognition, expert human curators evaluated 100 randomly chosen annotations each for cells, cytokines and diseases and found the automatic annotation to be 91%, 96% and 93% correct, respectively (**Supplementary Tables 6–8** and Online Methods). Similarly, for precision of relation extraction, we randomly chose 590 interaction evidence records (i.e., particular sentence instances) and manually evaluated entity recognition, ontology mapping, verb and relation detection, directionality, polarity and cellular function identification (**Supplementary Tables 9 and 10** and Online Methods). We observed a conservative true-positive rate of 75% when all of the metrics were considered, 82% when assessing triple precision (cell–cytokine–directionality) and 93% when checking cell–cytokine relation pair extraction only, ignoring directionality, polarity and other relation characteristics (**Fig. 1b**).

To evaluate performance in interaction recall (i.e., identifying known interactions), we created a gold-standard set by manually curating directional interactions from a reference book spanning up-to-date knowledge of cytokines²⁹ (Online Methods). Our machine learning-derived knowledgebase covered 79% of the interactions described in the reference book, yet it was approximately five times larger, containing an additional 3,055 directional interactions. Manual assessment of 200 of these yielded an 11.5% false-positive rate, suggesting that a large amount of biologically meaningful interactions appeared in only the primary literature.

Finally, we unified all identified interactions (manual and machine derived) into a single knowledgebase, which we named immuneXpresso (iX) (**Fig. 1c**). To quantify the advantages of this semantic-based approach, we compared its precision and recall with an alternative of assuming cells and cytokines co-occurring in the same sentence interact, without sentence structure analysis (Online Methods). Though the full set of co-occurrence-based relations showed 98% recall of both the reference book-curated and iX cell–cytokine pairs, 75.6% of these relations appeared in neither resource, suggesting a very high false-positive rate for a

co-occurrence approach, even when we set a threshold by a minimal number of repeat co-occurrences (**Fig. 1d**).

At present, iX contains a total of 4,118 directional cell–cytokine interactions (**Supplementary Table 11**), three times as many incoming interactions as outgoing ones, an enrichment qualitatively echoed in reference book-annotated interactions. These interactions stemmed from >31,000 articles (**Supplementary Table 12**). In addition, using the iX pipeline we collected annotations of thousands of diseases (11,260 distinct disease terms; 2,179 of them appearing in at least 100 papers) and identified mentions of 1,300 cell types (360 of which were hematopoietic) and 170 cytokines in these disease contexts (**Supplementary Tables 11 and 13**). iX is freely accessible for querying through <http://www.immunexpresso.org>, as well as via the ImmPort web site³⁰.

System-level characterization of intercellular interactions

iX offers an opportunity for a system-level view of intercellular information flow. Given the large number of cells and cytokines, many of which are poorly understood, we first grouped the cells into 16 major categories on the basis of cell ontology hierarchy and the cytokines into families on the basis of structure and function (Online Methods). This yielded a bipartite intercellular interaction network that showed information flow between cell types and cytokine families (**Fig. 2a**). We noted that cell types, irrespective of the number of identified cell subsets in their lineage, interacted with a large number of cytokine families. By replotting the network using the highest cell and cytokine resolution in iX, we observed an increase in distinct cell subset–cytokine profiles per cell type, particularly for outgoing interactions (**Supplementary Fig. 1a**, shown for CD4⁺ T cells). Yet, the bulk of interactions was still described solely at a low cellular resolution, which indicated that the unique cytokine milieu profile of distinct cellular subsets was, for the most part, still lacking.

Signaling of some cytokines may be highly specific or may broadly affect multiple cell subsets, constituting hubs of the intercellular interaction network. iX enables the study of global properties of the intercellular interaction network. For each interaction, we identified the highest cellular resolution it was reported for and, for each cytokine, calculated the number of cellular interactions it was associated with (i.e., its degree), covering both hematopoietic cell types (HPCs) and nonhematopoietic cell types. This demonstrated the existence of only a few hubs in the network, followed by a long tail of modestly and low-interacting cytokines (see **Fig. 2b** for incoming interactions and **Supplementary Fig. 1b** for outgoing interactions). We noted that 50% of the incoming interactions in the network were formed by 23 (16% of total) cytokines. These included the top hubs tumor necrosis factor (TNF), transforming growth factor (TGF)- β , interleukin (IL)-6 and interferon (IFN)- γ . Similarly, in the reverse outgoing direction, we attributed 50% of edges to 17 (15% of total) cytokines. Cytokine degrees in incoming and outgoing directions showed high correlation (**Fig. 2c**; Pearson’s $r = 0.086$), as in the gold-standard reference book (**Supplementary Fig. 1c**; Pearson’s $r = 0.69$). This correlation was lower, yet still observed, after removal of autocrine interactions that may have inflated similarity (Pearson’s $r = 0.73$ and $r = 0.36$ for iX and the reference book, respectively), as well as following removal of low-degree cytokines (Pearson’s $r = 0.5$), suggesting that hubs in the literature-derived intercellular interaction network appeared to be bidirectional, both targeting and secreted by a large number of cell types.

Immune intercellular network knowledge is biased

Analyzing the cytokine degree distribution, we could not reject a power-law distribution for either incoming or outgoing interactions

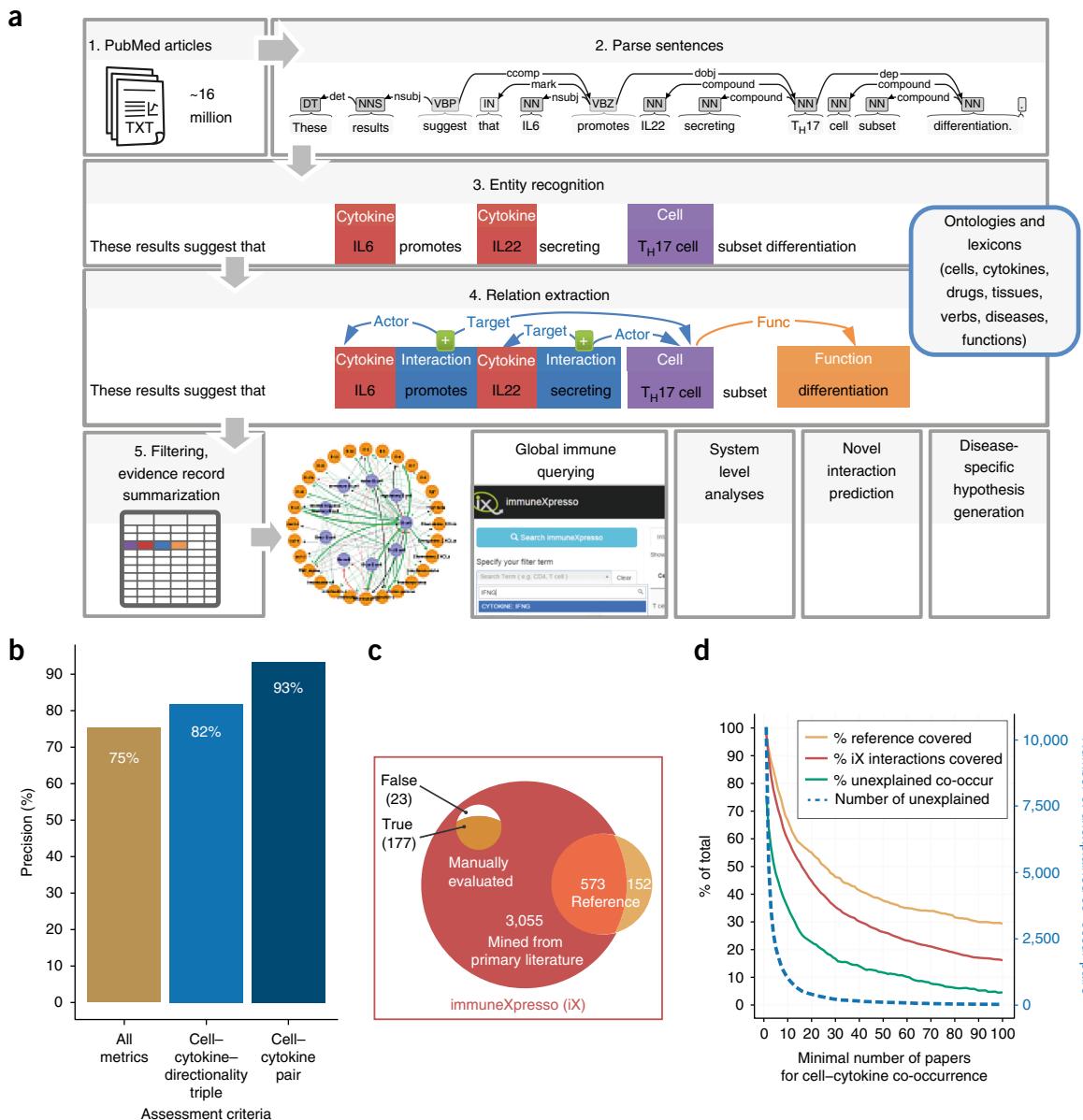


Figure 1 immuneXpresso assembles a system-level directional intercellular interaction network. (a) Schematic showing that PubMed abstracts were mined to identify cell, cytokine and context entities and map them to ontologies. Semantically linked cells, cytokines and verbs were extracted and characterized to assign interaction directionality, polarity and the biological function this interaction yielded. Results were scored and filtered to yield a global machine-readable view of immune intercellular signaling across a large breadth of conditions. (b) Evaluation of cell–cytokine relation extraction precision. 590 randomly selected evidence records (i.e., relations extracted from individual sentences) were evaluated by a human curator to assess entity recognition performance and cell–cytokine relation extraction (including detection of verb, directionality, polarity and, where relevant, biological function). (c) Venn diagram depicting the precision (88.5%) of literature-derived cell–cytokine interactions (i.e., cell–cytokine–directionality triples) and coverage (79%) of the manually annotated reference book²⁹. (d) Evaluation of precision and recall of co-occurrence–based relation extraction. Cell–cytokine pairs, linked by co-occurrence within sentence boundaries, were filtered by their evidence strength (x axis). Coverage of relations derived from the reference book and from the semantic-parsing approach (“iX interactions”) is shown, as well as the percentage (or amount as a dotted line, right y axis) of co-occurring pairs that appeared in none of the resources above, which likely reflects the false-positive rate.

(Supplementary Fig. 2, $P = 0.73$ and $P = 0.47$, respectively, for incoming and outgoing interactions; Online Methods). Heavy-tailed network distributions may arise due to a research bias, yielding a ‘rich-get-richer’ phenomenon^{31,32}. Conversely, such degree distributions may arise naturally due to biological network structure³³. Analysis of cytokine interaction knowledge accumulation showed the existence of

one or two connection-rich leaders per cytokine family, with the other family members trailing well behind (Fig. 3a and Supplementary Fig. 3). These were predominantly founding cytokine family members, such as IL-6, IFN- γ , IL-10 and TNF, and they maintained their overwhelming dominance even when we discarded all explicit references to global family mentions in the text (e.g., TNF family). We

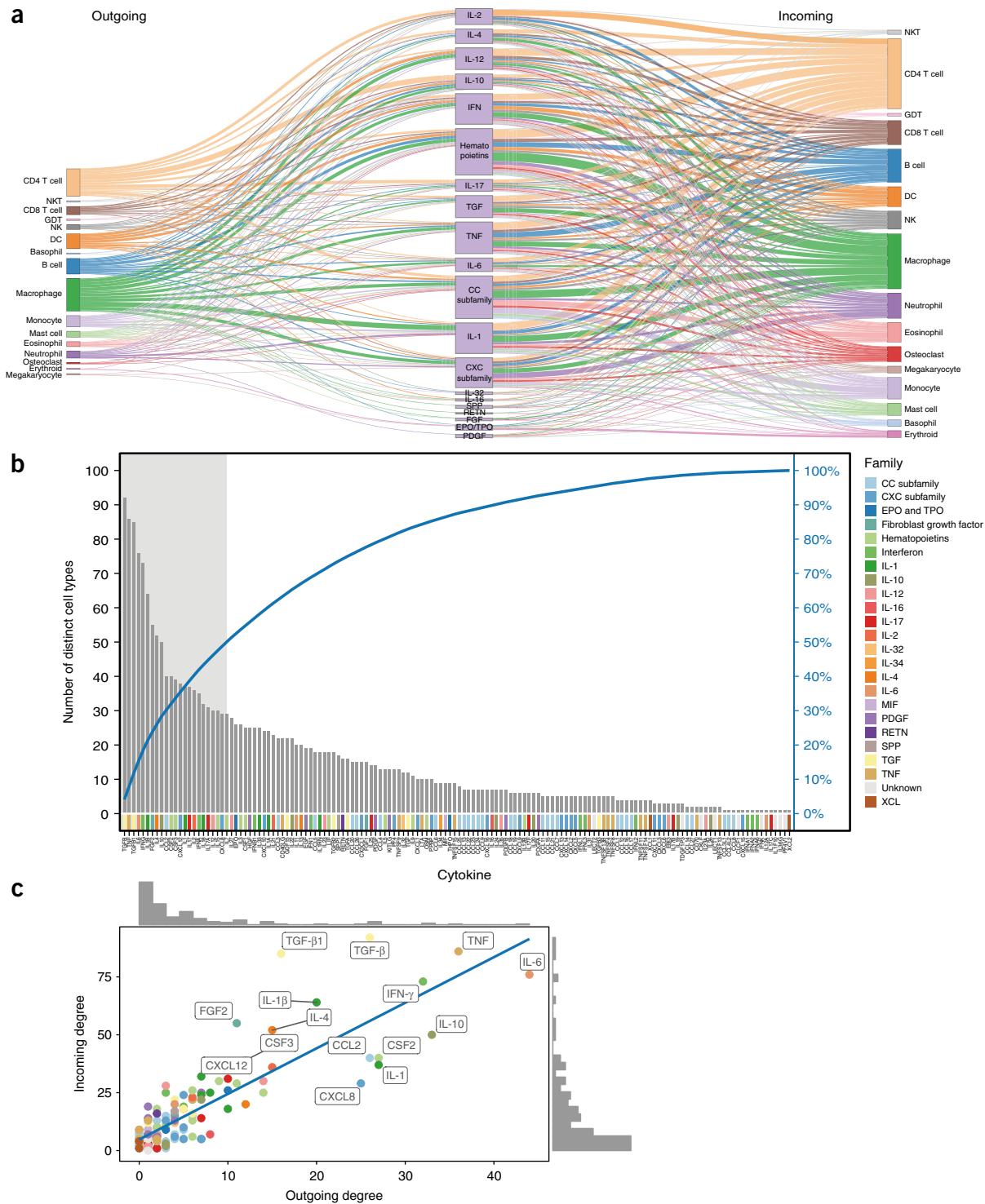


Figure 2 System-level characteristic of intercellular information flow in the literature-derived network. (a) Sankey plots showing bipartite information flow between cellular secretion of cytokine families and those cytokine families that affect a variety of cell subsets. (b) A sorted histogram illustrating the number of unique cell types affected by each of the 144 cytokines (incoming interactions). Second y axis displays the information as a cumulative sum (blue line). 50% of incoming edges are attributed to only 23 (16%) cytokines (grey area). Cytokine family classification appears as coloring of individual members along the x axis. (c) Scatter plot highlighting the strong correlation in cytokine degrees between incoming and outgoing directions ($n = 145$ cytokines, Pearson's $r = 0.86$, $P < 0.001$).

detected a low global correlation between a cytokine's date of discovery and its degree, suggesting that intercellular communication knowledge had not reached saturation, either for hubs or for less-connected

cytokines (Supplementary Fig. 4; Pearson's $r = -0.27$ and $r = -0.26$ for incoming and outgoing interactions, respectively, which were driven by a few highly dominant hubs). Analysis of recent 5-year change in

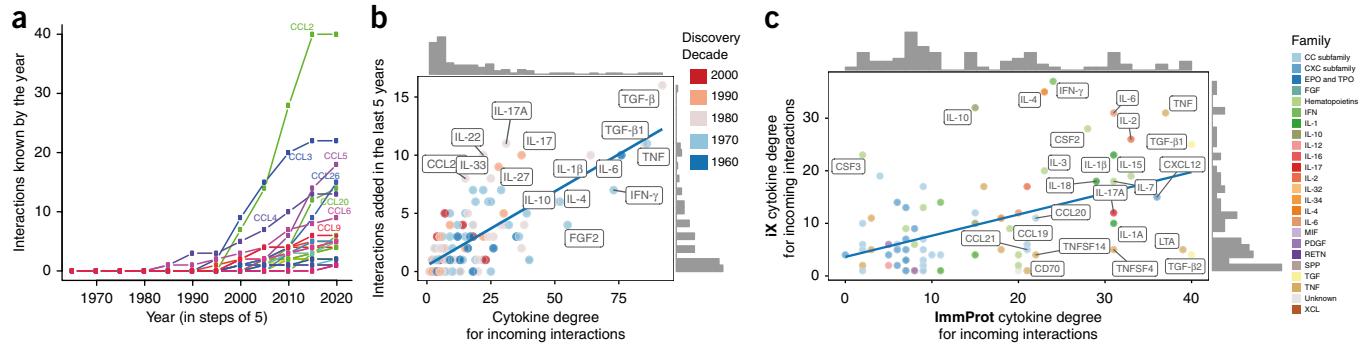


Figure 3 Immune intercellular network knowledge is biased and far from saturated. **(a)** Unequal information gain within cytokine families. Representative plot using the CC chemokine subfamily of the incoming interaction knowledge gain over time, in a 5-year window from initial date of the chemokine's identification in the text. Colored lines represent individual chemokine family members and display the total number of distinct cell types known to interact with the chemokine by the respective year (*x* axis). **(b)** The number of distinct cell types a cytokine is known to act on (*x* axis) is positively correlated with the number of new cytokine–cell interactions added in recent years (*y* axis; $n = 134$ cytokines, Pearson's $r = 0.76$, $P < 0.001$), yet up-and-coming hubs, such as IL-22, are identifiable as well. **(c)** Cytokine degrees in the iX knowledgebase (incoming interactions, restricted to hematopoietic cells) are positively correlated with those derived from the characterization of cytokine receptors on cells in the ImmProt project³⁴ ($n = 82$ cytokines present in both data sets, Spearman's $\rho = 0.38$, $P < 0.001$). Cytokines below the line suggest putative currently unknown interactions.

connectivity degree suggested that for some hubs, such as FGF2, their iX degree likely reflected cellular interactivity potential, whereas others, such as IL-6 and TGF- β , were still accumulating new connections at a high rate (Fig. 3b).

To assess how well the literature-derived knowledge represented experimental data, we compared iX cytokine degrees to those obtained from ImmProt, a high-resolution proteomics compendium quantifying protein expression in resting and activated states in more than 20 sorted immune cell types³⁴. We approximated outgoing and incoming interactions based on the expression of cytokine and cytokine receptor proteins in ImmProt cellular profiles. Comparison of incoming cytokine degrees in the resulting experimental network with the literature-derived ones (Fig. 3c) demonstrated significant correlation (Spearman $\rho = 0.38$; $P < 0.001$; based on 82 cytokines present in both data sets), with lower correlation for the opposite direction (Spearman $\rho = 0.26$; $P = 0.1$; based on 41 shared cytokines), likely due to the fact that the ImmProt secretion profiles were measured in a single condition.

Prediction of cell–cytokine interactions

Using the subset of the iX knowledgebase comprised of articles published up to 2014, we systematically predicted cell–cytokine interactions, using three orthogonal approaches (Fig. 4a and Online Methods): (i) an unsupervised analysis of iX interaction profiles, (ii) a supervised analysis of iX cytokine–cell interactions grouped by cytokine families, as well as by (iii) contrasting iX information to receptor or cytokine expression data, as reflected in the ImmProt³⁴ and ImmGen databases³⁵ (e.g., LTA in Fig. 3c). This systematic prediction process yielded 472 incoming and 367 outgoing ranked interaction candidates (Supplementary Tables 14 and 15). Of these, we manually evaluated 78 predictions by extensive literature searches (Fig. 4b). This process identified 55% of the candidates as already observed, true-positive interactions, 3% with evidence published later than the 2014 training set, and 3% with evidence of cytokine receptor expression only. This high rate of recovery of known interactions was reassuring and suggested that the remaining 40% of candidate interactions were enriched for previously unrecognized interactions.

We tested the validity of two top-rated candidate interaction predictions: for IL-7, our supervised cytokine family–prediction approach

suggested an outgoing interaction (i.e., secretion) in monocytes (the literature currently describes the opposite interaction only, i.e., that IL-7 affects monocytes³⁶). In agreement with the prediction, we observed IL-7 production by monocytes in the activated human peripheral blood mononuclear cell (PBMC) population. In addition, as expected, we detected dendritic cell production of IL-7, and no IL-7 production in CD8⁺ T cells, as reported in the literature (Fig. 4c). Similarly, our unsupervised prediction approach suggested that IL-34 activates signaling in T cells. We also detected expression of a corresponding receptor, CSF1R, on resting CD8⁺ T cell subsets in ImmProt³⁴ (Supplementary Fig. 5). We stimulated human PBMCs with IL-34 and observed robust phosphorylation of the kinase ERK in monocytes (Fig. 4d), as has previously been reported³⁷. We also observed activation of the transcription factors NF- κ B and STAT5 by IL-34 in CD8⁺ effector memory T cells (Fig. 4d). Moreover, our prediction also supported our recent validation of CD4⁺ memory cell induction by IL-34 signaling following upregulation of CSF1R during activation³⁴.

Immune-centric classification of diseases

We reasoned that the structured format and breadth of iX could be leveraged to obtain an immune-centered perspective on diseases and their relations. To do so, we picked a set of 188 broadly studied diseases whose associated abstracts we sampled to obtain a characteristic intercellular immune profile (Fig. 5a, Supplementary Fig. 6 and Supplementary Table 16). We clustered these in an unsupervised manner to assemble an immune-centered map of disease similarities and differences (Online Methods).

Analysis of this clustering outcome divided diseases into 18 modules on the basis of associated cells, cytokines and interactions (Fig. 5b). We observed mixed agreement of this classification with a clinically based one (SNOMED)—some modules clustered clinically similar phenotypes; for example, module 2 showed strong clustering of cardiovascular diseases, whereas module 9 captured inflammatory bowel disorders and psoriasis. Similarly, cancers were grouped into four modules on the basis of tissue type. In contrast, in some cases the immune-centered clustering yielded modules that included diseases from presumably unrelated clinical conditions—modules 14 and 15 suggested a high degree of immune similarity between metabolic

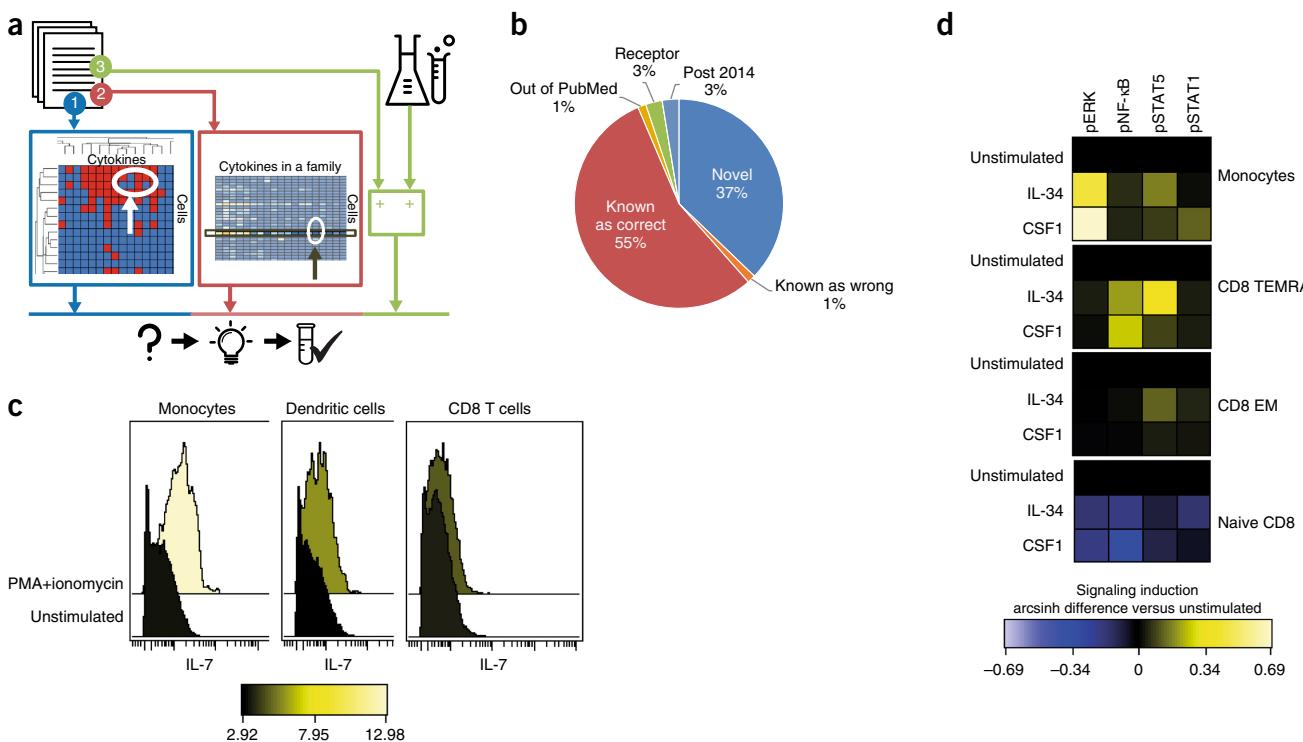


Figure 4 Prediction and validation of cell–cytokine interactions. **(a)** Diagram illustrating three different methodologies for predicting cell–cytokine interactions: (1) by unsupervised clustering and filling in of missing ‘gaps’, (2) through supervised analysis of cytokine families and (3) by comparison to measured external mRNA or proteome data. **(b)** Manual evaluation of the predicted high-confidence interactions using free literature searches. **(c)** IL-7 production by monocyte, dendritic cell and CD8 T cell subsets of freshly isolated PBMCs from two healthy donors (PMA+ionomycin, 4 h) by intracellular cytokine staining (CyTOF). **(d)** Heat map summary of NF-κB and STAT5 phosphorylation in CD8+ cell subsets after IL-34 stimulation (100 ng/ml, 15 min) of freshly isolated human PBMCs. EM, effector memory CD8 T cells (CCR7-CD45RA-); Temra, CD45RA+ effector memory CD8 T cells (CCR7-CD45RA+).

disorders and a subset of cardiovascular diseases, an observation that was also supported by high interconnectivity of modules 14 and 15 with module 2.

We used iX to automatically generate an intercellular immune-interaction map for type 2 diabetes on the basis of 5,484 abstracts (Fig. 5c and Online Methods). The network recapitulated the molecular basis of the disease and captured the key role of tissue-resident inflammatory macrophages, monocytes, fat cells and preadipocytes in secreting the pro-inflammatory cytokines TNF, IL-6 and IL-1 β , which trigger the release of adipokines that contribute to development of insulin resistance³⁸. This profile co-clustered with other diseases in module 15, which consist of metabolic and cardiovascular diseases, an appreciated link³⁹. Enrichment analysis of module 15 and its closely associated module 2, as compared to all others (Fig. 5d and Online Methods), showed a strong association with the pro-inflammatory adipokine RETN and its interaction with monocytes. RETN has been proposed to link the heightened inflammatory state in aged and obese individuals to insulin resistance, vascular inflammation and low-density lipoprotein (LDL)-cholesterol levels, thus contributing to the risk of developing metabolic syndrome⁴⁰. Moreover, elevated levels of RETN directly induce IL-6 secretion⁴¹, which contributes to the constant low-grade inflammatory state associated with age and cardiovascular conditions^{42,43}. Thus, whereas modules 15 and 2 phenotypically represent different pathological conditions, their molecular commonalities suggest consideration for common therapeutic interventions.

Prediction of cytokine–disease associations

Analysis of cytokine profiles across the 188 diseases revealed three predominant cytokine classes with respect to disease: disease specific, module specific and backbone cytokines that were associated with the overwhelming majority of tested diseases (Fig. 6a, Supplementary Table 17 and Online Methods). We noted a high overlap between backbone cytokines and those previously identified as hubs (Fig. 2b), suggesting that their pan-disease universality stems from the dominant role in the overall intercellular interaction network.

We hypothesized that given the unequal within-module knowledge per disease we might be able to predict *de novo* cytokine–disease associations. For each module, we hierarchically clustered disease subsets across all cytokines and systematically predicted cytokine–disease associations (Online Methods). This yielded over 466 ranked candidates, which stemmed from all 18 modules (Supplementary Table 18). As we assembled the disease immune profiles by sampling, we checked the predicted cytokine–disease associations on the full, nonsampled iX knowledgebase, which validated 81% of predicted interactions as true positives and suggested a likely enrichment for unappreciated cytokine–disease associations in the remaining set (Supplementary Fig. 7).

To test this, we looked for experimental confirmation of two of the top-rated candidate associations, the chemokines CCL8 and CCL24 in psoriasis, which, to the best of our knowledge, have not been reported. We analyzed two publicly available gene expression data sets for psoriasis^{44,45} and found CCL8 to be significantly upregulated in psoriatic

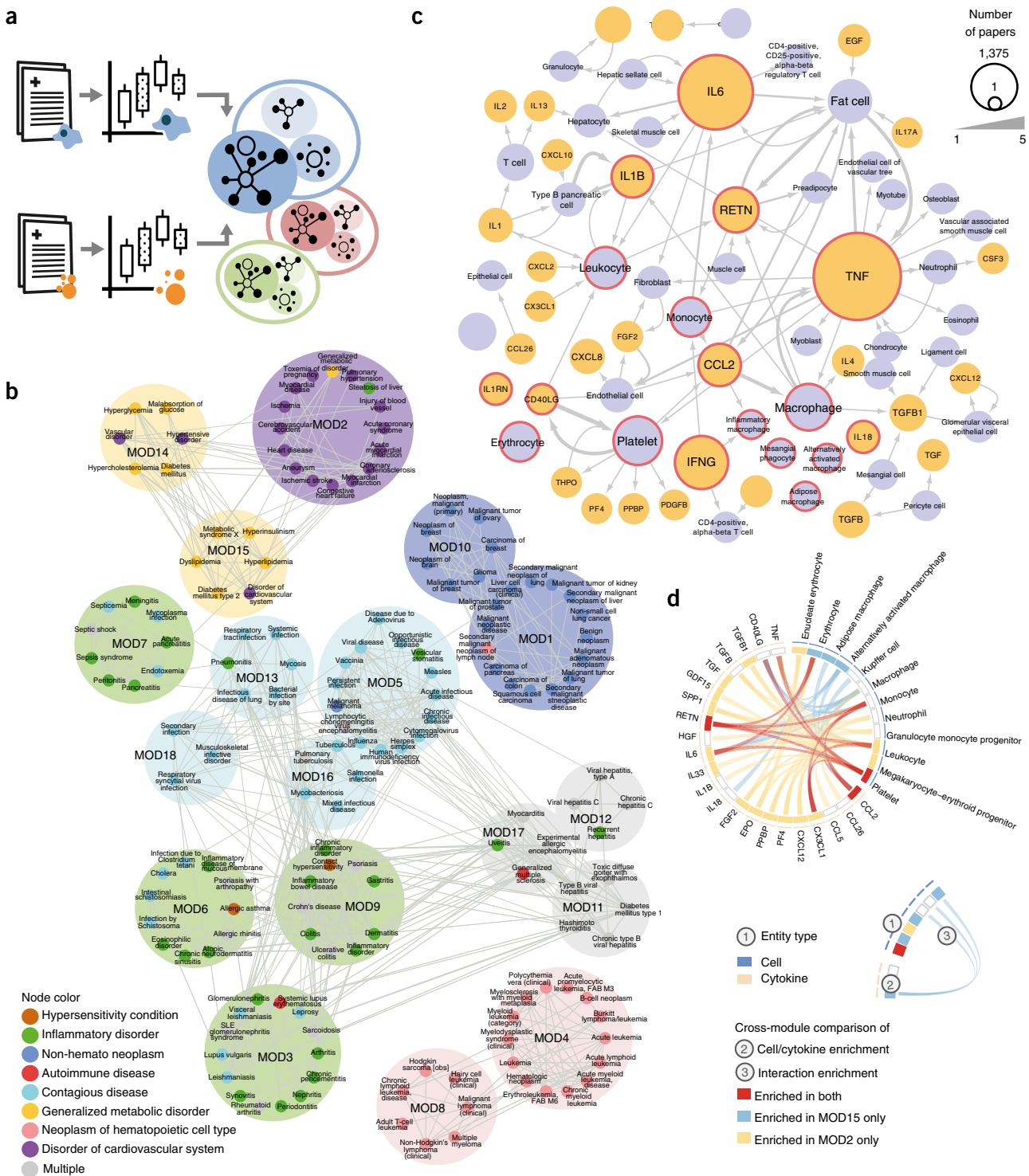


Figure 5 Global interaction profile analysis across diseases. **(a)** Diagram illustrating the methodology of disease similarity module extraction: for 188 top-cited diseases, papers co-mentioning the disease either with cytokines or with hematopoietic cells were repeatedly sampled to identify cells and cytokines mentioned frequently in the particular context. Those entities, together with their potential interactions (as reported in the overall iX knowledgebase), formed disease-signaling profiles that were used to build a global disease similarity map. **(b)** The immune-centered global map of disease similarities and differences, assembled in an unsupervised manner for 188 top-cited diseases. Nodes in the graph correspond to diseases (colored by their known clinical disease categories); edges correspond to significant correlation between diseases. Modules were colored by the most abundant SNOMED category and demonstrated a large overlap between module separation and clinical classification. **(c)** iX extracts context-specific cells (purple), cytokines (orange) and interactions for a variety of conditions (here diabetes mellitus type 2, as reported in the overall iX knowledgebase), providing a global view of the condition-specific immune signaling components and emphasizing its key players. Cells and cytokines forming the sampling-based disease immune profile are in red. **(d)** Circos plot showing enriched module signatures (cell, cytokine and interactions) relative to the background for cardiovascular and metabolic syndrome diseases (modules 2 and 15).

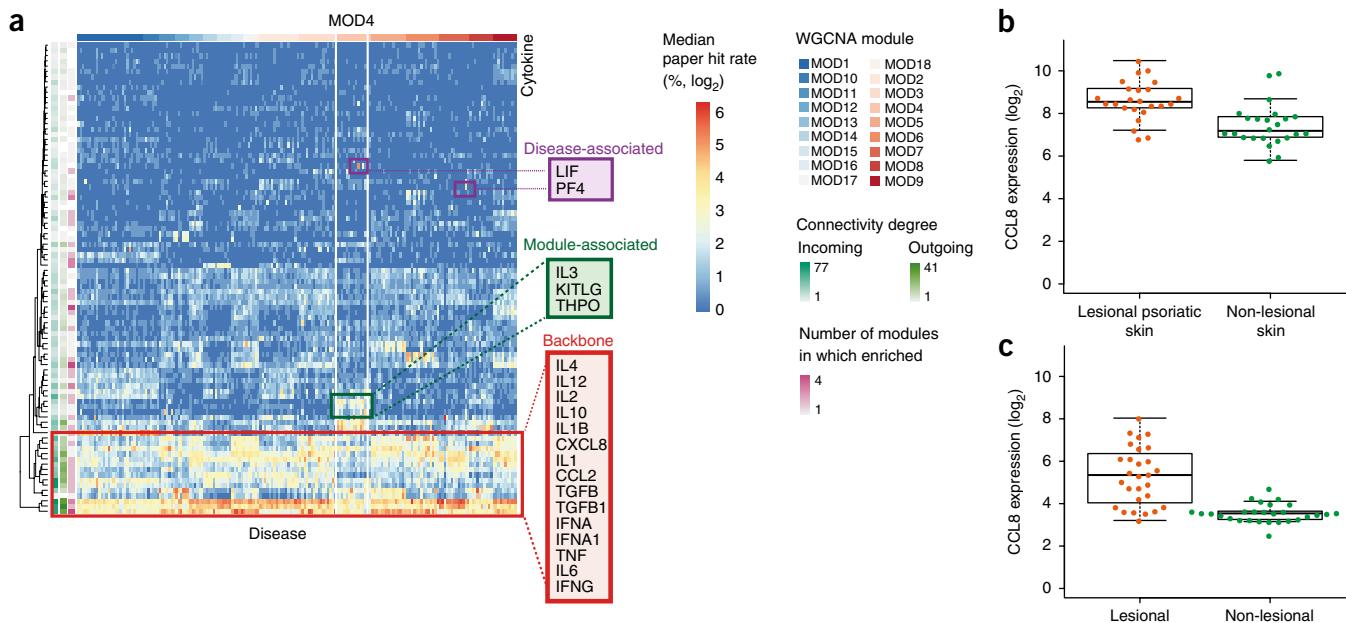


Figure 6 iX informs less well-studied diseases. **(a)** Heat map showing frequency of cytokine mentions across articles in the context of the 188 top-cited diseases. Color scale indicates the median percentage (\log_2 -transformed) of paper hits for each cytokine (rows) in the context of each disease (columns); sorted by the unsupervised module classification across 200 paper-sampling iterations. An unequal immune system utilization is observed with respect to three cytokine classes: backbone cytokines highly observed across most modules, those that are module specific and those specifically enriched in select diseases. Class examples are highlighted in red. Annotation tracks on the left indicate cytokine degrees. **(b,c)** CCL8 is differentially expressed in psoriasis. The expression of CCL8 in healthy skin versus lesion biopsies of psoriatic patients at baseline is shown (paired two-sided Wilcoxon signed-rank test; $P = 2.47 \times 10^{-5}$, $n = 24$ independent samples **(b)**; $P = 1.04 \times 10^{-7}$, $n = 28$ independent samples **(c)**). Box plot elements: center line, median; box limits, first to third quartile (Q1 to Q3); whiskers, extend to the most extreme data point within 1.5x the interquartile range (IQR) from the box; circles, data points. Validation of the predicted chemokine–disease association in two independent public gene expression data sets^{44,45}.

lesions versus healthy skin in both data sets (paired two-sided Wilcoxon signed-rank test, $P = 0.0048$, $P = 2.47 \times 10^{-5}$; **Fig. 6b,c**). For CCL24 we observed significant differences in one study⁴⁵ ($P = 1.04 \times 10^{-7}$).

DISCUSSION

Knowledge of the immune intercellular network is crucial for understanding immune responses in health and disease. However, the high system complexity leaves even expert researchers struggling to maintain a mental picture of the immune milieu and often leads to knowledge biases. We leveraged a computational model of intercellular interaction network knowledge to accelerate discovery of interactions and the context in which they occurred, identify disease-associated immune profiles and build an immune-centric disease classification partially overlapping with existing clinical disease classification.

The knowledge we captured was unique in breadth and resolution, yet it could be expanded even further. Full-text article support, as well as identification of interactions described beyond sentence boundaries and using more complex sentence structures, can boost the volume of captured evidence and our ability to obtain more reliable and richer view of interactions, with less bias, particularly of those interactions that are understudied. We foresee extension of this approach to also include additional events, such as direct cell–cell interactions and downstream intercellular signaling to capture complex cellular interaction cascades. The extensive metadata we extracted for each article, including MESH terms and bibliographic information, together with detailed characterization of the captured interactions, could be used for advanced filtering, which would allow focus on the most authoritative knowledge. Beyond this, we envision that the structured formatting

of knowledge we have achieved can be leveraged by machine-learning applications, using statistical analyses of domain frequency and chronological pattern biases to identify potential discrepancies and erroneous claims in the published knowledge.

Technological advances now enable us to step beyond assaying a narrow set of measures to high dimensional phenotyping across the breadth of the immune system at an unprecedented scale. These data are primarily analyzed by statistical analysis methods that are geared to identify differences and correlations yet that lack any back-end model of the immune system’s structure and function, lack in their ability to leverage domain knowledge or interpret what these differences mean. This results in an interpretation process that is primarily manual, not systematized, and that relies strongly on investigator conjecture. Intelligent systematized interpretation requires having a machine-readable map of how immune components are connected and a formalized reasoning framework on which one could test hypotheses and refine knowledge. Here we built immuneXpresso, a framework that structures and standardizes our knowledge of immune intercellular interactions, under many conditions, and updates periodically. Its integration with high-dimensional immune data will enable paradigms of reasoning over heterogeneous cell populations, making first steps towards transforming immunology to systemized, model-based science, a true ‘systems immunology’.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank A. Butte and M. Davis for fruitful discussions and advice, N. Geifman for assistance with cytokine ontology development, D. Dougall for contribution to the cell lexicon, members of the Shen-Orr lab for reference book curation, D. Cohen for the high-performance computing cluster support, R. Rechart for Text Mining insights, and P. Dunn and S. Bhattacharya for the user interface development support. This work was supported by the US National Institutes of Health (NIH)-National Institute of Allergy and Infectious Diseases (U19 AI057229, and BISC contract HHSN272201200028C) and an award from the Rappaport Family Institute for Research in the Medical Sciences (S.S.-O.).

AUTHOR CONTRIBUTIONS

K.K. designed, performed and interpreted the analyses, led the design and development of the software, implemented relation extraction, filtering and network assembly, and wrote the manuscript; E.S. interpreted the analyses, performed quality control, designed and performed the experimental validations, and wrote the manuscript; A.Z.-K. implemented indexing, ontology support and reference book annotation software, and broadly contributed to the entire computational pipeline development; Y.K. conceived the pipeline architecture and broadly contributed to the software formation; Y.G. assisted with cell entity recognition; G.S.-M. performed quality control on Text Mining output and assisted with cytokine ontology development; T.D. assisted with quality control on Text Mining output; M.B. contributed to quality control and prediction evaluation; N.A.-R. wrote the software and implemented the website back end; J.C. and J.W. designed and developed the user interface; J.C.R. and F.M. provided and interpreted the proteomic data; N.A. assisted with machine-learning for quality control; D.R. contributed to quality control and interpretation of disease profiles; and S.S.S.-O. conceived the idea, oversaw, designed and interpreted the analyses, and wrote the manuscript.

COMPETING INTERESTS

K.K. and Y.K. are employees and co-founders of CytoReason. S.S.S.-O. and E.S. are co-founders of, and serve as scientific advisors and/or consultants to, CytoReason.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Maecker, H.T. *et al.* New tools for classification and monitoring of autoimmune diseases. *Nat. Rev. Rheumatol.* **8**, 317–328 (2012).
2. Hoffmann, R. & Valencia, A. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics* **21** (Suppl. 2), ii252–ii258 (2005).
3. Jimeno, A. *et al.* Assessment of disease-named-entity recognition on a corpus of annotated sentences. *BMC Bioinformatics* **9** (Suppl. 3), S3 (2008).
4. Leaman, R., Islamaj Dogan, R. & Lu, Z. DNORM: disease name normalization with pairwise learning to rank. *Bioinformatics* **29**, 2909–2917 (2013).
5. McDonald, R.T. *et al.* An entity tagger for recognizing acquired genomic variations in cancer literature. *Bioinformatics* **20**, 3249–3251 (2004).
6. Tanenblatt, M., Coden, A. & Sominsky, I. The ConceptMapper approach to named-entity recognition. *LREC Int. Conf. Lang. Resour. Eval.* **546–551** (2010).
7. Funk, C. *et al.* Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. *BMC Bioinformatics* **15**, 59 (2014).
8. Shah, N.H. *et al.* Comparison of concept recognizers for building the Open Biomedical Annotator. *BMC Bioinformatics* **10** (Suppl. 9), S14 (2009).
9. Bada, M. *et al.* Concept annotation in the CRAFT corpus. *BMC Bioinformatics* **13**, 161 (2012).
10. Kim, J.D., Ohta, T., Tateisi, Y. & Tsujii, J. GENIA corpus—semantically annotated corpus for bio-textmining. *Bioinformatics* **19** (Suppl. 1), i180–i182 (2003).
11. Arighi, C.N. *et al.* Overview of the BioCreative III Workshop. *BMC Bioinformatics* **12** (Suppl. 8), S1 (2011).
12. Kim, J., Ohta, T., Pyysalo, S. & Kano, Y. Overview of BioNLP 2009 shared task on event extraction. (Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task) 1–9 (Association for Computational Linguistics, Stroudsburg, Pennsylvania, USA, 2009).
13. Kim, J.-D. *et al.* Overview of BioNLP Shared Task 2011. (Proceedings of the BioNLP Shared Task 2011 Workshop) 1–6 (Association for Computational Linguistics, Stroudsburg, Pennsylvania, USA, 2011).
14. Kim, J.D., Ohta, T. & Tsujii, J. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics* **9**, 10 (2008).
15. Ananiadou, S., Pyysalo, S., Tsujii, J. & Kell, D.B. Event extraction for systems biology by text-mining the literature. *Trends Biotechnol.* **28**, 381–390 (2010).
16. Pyysalo, S. *et al.* Event extraction across multiple levels of biological organization. *Bioinformatics* **28**, i575–i581 (2012).
17. Mahmood, A.S.M.A., Wu, T.J., Mazumder, R. & Vijay-Shanker, K. DiMeX: a text-mining system for mutation-disease association extraction. *PLoS One* **11**, e0152725 (2016).
18. Lee, K. *et al.* BRONCO: Biomedical entity Relation ONcology CORpus for extracting gene-variant-disease-drug relations. *Database (Oxford)* **2016**, 1–13 (2016).
19. Verspoor, K.M., Heo, G.E., Kang, K.Y. & Song, M. Establishing a baseline for literature-mining human genetic variants and their relationships to disease cohorts. *BMC Med. Inform. Decis. Mak.* **16** (Suppl. 1), 68 (2016).
20. Liu, H., Hunter, L., Kešelj, V. & Verspoor, K. Approximate subgraph matching-based literature mining for biomedical events and relations. *PLoS One* **8**, e60954 (2013).
21. Björne, J. *et al.* Extracting complex biological events with rich graph-based feature sets. (Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task 10–18 (Association for Computational Linguistics, Stroudsburg, Pennsylvania, USA, 2009).
22. Rzhetsky, A., Seringhaus, M. & Gerstein, M.B. Getting started in text mining: part two. *PLoS Comput. Biol.* **5**, e1000411 (2009).
23. Zhu, F. *et al.* Biomedical text mining and its applications in cancer research. *J. Biomed. Inform.* **46**, 200–211 (2013).
24. Jensen, L.J., Saric, J. & Bork, P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.* **7**, 119–129 (2006).
25. Goh, K.-I. *et al.* The human disease network. *Proc. Natl. Acad. Sci. USA* **104**, 8685–8690 (2007).
26. Linghu, B., Snitkin, E.S., Hu, Z., Xia, Y. & Delisi, C. Genome-wide prioritization of disease genes and identification of disease–disease associations from an integrated human functional linkage network. *Genome Biol.* **10**, R91 (2009).
27. Hu, G. & Agarwal, P. Human disease–drug network based on genomic expression profiles. *PLoS One* **4**, e6536 (2009).
28. Kilpinen, S. *et al.* Systematic bioinformatic analysis of expression levels of 17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues. *Genome Biol.* **9**, R139 (2008).
29. Dembic, Z. *The Cytokines of the Immune System: The Role of Cytokines in Disease Related to Immune Response* (Elsevier Science, 2015).
30. Bhattacharya, S. *et al.* ImmPort: disseminating data to the public for the future of immunology. *Immunol. Res.* **58**, 234–239 (2014).
31. Edwards, A.M. *et al.* Too many roads not taken. *Nature* **470**, 163–165 (2011).
32. Barabasi, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
33. Barabasi, A.-L. Scale-free networks: a decade and beyond. *Science* **325**, 412–413 (2009).
34. Rieckmann, J.C. *et al.* Social network architecture of human immune cells unveiled by quantitative proteomics. *Nat. Immunol.* **18**, 583–593 (2017).
35. Heng, T.S.P. & Painter, M.W. The Immunological Genome Project: networks of gene expression in immune cells. *Nat. Immunol.* **9**, 1091–1094 (2008).
36. Möller, P., Böhm, M., Czarnetzki, B.M. & Schadendorf, D. Interleukin-7: Biology and implications for dermatology. *Exp. Dermatol.* **5**, 129–137 (1996).
37. Lin, H. *et al.* Discovery of a cytokine and its receptor by functional screening of the extracellular proteome. *Science* **320**, 807–811 (2008).
38. Asghar, A. & Sheikh, N. Role of immune cells in obesity-induced low-grade inflammation and insulin resistance. *Cell. Immunol.* **315**, 18–26 (2017).
39. Gaede, P. *et al.* Multifactorial intervention and cardiovascular disease in patients with type 2 diabetes. *N. Engl. J. Med.* **348**, 383–393 (2003).
40. Park, H.K., Kwak, M.K., Kim, H.J. & Ahima, R.S. Linking resistin, inflammation and cardiometabolic diseases. *Korean J. Intern. Med.* **32**, 239–247 (2017).
41. Hillenbrand, A., Weiss, M., Knippelschild, U., Wolf, A.M. & Huber-Lang, M. Sepsis-induced adipokine change with regard to insulin resistance. *Int. J. Inflam.* **2012**, 972368 (2012).
42. Shen-Orr, S.S. *et al.* Defective signaling in the JAK–STAT pathway tracks with chronic inflammation and cardiovascular risk in aging humans. *Cell Syst.* **3**, 374–384.e4 (2016).
43. Furman, D. *et al.* Expression of specific inflammasome gene modules stratifies older individuals into two extreme clinical and immunological states. *Nat. Med.* **23**, 174–184 (2017).
44. Russell, C.B. *et al.* Gene expression profiles normalized in psoriatic skin by treatment with brodalumab, a human anti-IL-17 receptor monoclonal antibody. *J. Immunol.* **192**, 3828–3836 (2014).
45. Yao, Y. *et al.* Type I interferon: potential therapeutic target for psoriasis? *PLoS One* **3**, e2737 (2008).

ONLINE METHODS

iX pipeline execution environment. The computational pipeline to assemble the iX knowledgebase was executed on a high-performance computing cluster, by running a batch job scheduling system with up to 150 simultaneous jobs allowed. Details of the specific pipeline steps appear below. It typically took ~2 weeks to generate the iX database from start to finish for the entire corpus.

Corpus selection, parsing and indexing. Abstracts of all English-language articles published electronically between 1960 and July 2017 (~16 million) were downloaded from PubMed using the EFetch utility. For each abstract, we extracted additional metadata, such as the article title, year of publication, whether it was a review or not, and its assigned MESH terms. We focused on the mammalian immune system because of the rapid evolution of the immune system^{46,47}, and in particular restricted our analysis to abstracts that were assigned a mouse and/or human MESH term, as these were the prime organisms relevant for biomedical research for which information is available. We used the Stanford Parser engine⁴⁸ to split the abstract text to sentences and words, perform part-of-speech tagging, and extract sentence syntactic structure and grammatical relations (i.e., ‘typed dependencies’). This information was then indexed within the Elasticsearch engine (<https://www.elastic.co/>) to allow querying for sub-corpora that potentially contained biological entities of interest. Moreover, preserving text processing products in the index eliminated the need for future time-consuming abstract reparsing, if entity recognition was expanded to support additional entity types (e.g., drugs, tissues or pathways).

Entity recognition and ontologies. To identify mentions of biological entities of interest (diseases, cells and cytokines) within article abstracts, we applied a dictionary-based approach with dictionaries either adapted from standard public knowledgebases or assembled from scratch. For diseases, we first queried Elasticsearch for articles that contained synonym phrases, listed by manually curated compilation of UMLS Metathesaurus⁴⁹, and then post-processed the returned abstracts, sentence-by-sentence, to look for the matches and extract precise information for them. In particular, for each identified disease mention, we recorded its position in the sentence, the particular synonym used, as well as the public ‘concept unique identifier’ (CUI) and concept name, as defined by the SNOMED CT-controlled vocabulary of the Metathesaurus (Supplementary Tables 8 and 13). This post-processing stage dropped conditions contained within longer disease entities (e.g., ‘deficiency’ and ‘vitamin deficiency’ were dropped, and only the most specific ‘vitamin A deficiency’ term was retained within the sentence “The essential role of vitamin A in kidney development has been demonstrated in vitamin A deficiency and gene-targeting studies.”). To achieve this, we automatically examined all diseases with overlapping sentence positions to retain those with the longest position span, and, among them, containing the highest number of words.

For cells, initial testing suggested that straightforward lookup of terms contained within the official Cell Ontology⁵⁰ would miss a substantial fraction of cell occurrences in text due to the large number of possible forms of describing cell subsets. This pluralism in naming was hard to anticipate automatically, both when describing cells by name (e.g., ‘human CD8⁺ terminally differentiated memory cell’ does not appear in the Cell Ontology and would not be captured by straightforward dictionary lookup) or by cell surface marker combination (e.g., CD3⁺CD4⁺CD45RA⁺ cell), whose delineation in the Cell Ontology is limited. To resolve this, we expanded the Cell Ontology with a manually curated set of synonyms, and, more importantly, introduced a small lexicon of seed words that served as a starting point, an anchor, for cell recognition in sentences (e.g., cell, lymphocyte and macrophage; see Supplementary Table 1 for the full list).

The cytokine dictionary was assembled manually, due to lack of an established lexicon (Supplementary Note 1).

For cell and cytokine entity recognition, we queried Elasticsearch for articles containing either a mention of a cytokine synonym or a cell ‘seed’ word. The focus was on articles that mentioned either or both cell and/or cytokine, to serve the basis for further relation extraction. Akin to diseases, we post-processed candidate articles to better characterize the matches and, for cells, to expand from seeds to often multi-word hard-to-anticipate cell name phrases,

using typed dependencies (Supplementary Fig. 8 and Supplementary Note 2). In addition, the captured cell phrases were mapped to the Cell Ontology (Supplementary Note 3).

Lastly, following cell- and cytokine-mention candidate extraction, we analyzed their within-sentence positions to filter out erroneous identification of overlapping entities from different ontologies (e.g., erroneous ‘granulocyte’ or ‘macrophage’ cell matches within the cytokine entity ‘granulocyte–macrophage colony-stimulating factor’). For all remaining cell and cytokine mentions, we recorded their start and end positions in a sentence, the particular phrase or synonym used, as well as the representative ID and concept name, assigned either by official Cell Ontology for cells or by our manually constructed dictionary for cytokines (Supplementary Tables 6 and 7; see Supplementary Tables 1 and 3 for frequencies of terms captured for each cell seed and cytokine lexicon concept, respectively).

Relation extraction. Following extraction of articles containing either the cytokine synonym or cell seed word mentions, we applied sentence-by-sentence post-processing to detect verbs, and when possible, linked cells, cytokines and verbs into relations. We analyzed sentence typed dependencies⁴⁸ to detect semantically related cells and cytokines within sentence boundaries and identify the directionality (i.e., a cytokine acting on a cell, like ‘IL-6 promotes T_H17 cell differentiation’, or the opposite, a cell producing a cytokine, like ‘T cells secrete IL-2’), polarity (i.e., positive, negative or neutral effect of the interaction, such as upregulation, inhibition or just alteration of a cell function, respectively), as well as the cellular biological function impacted by this relation (e.g., cell proliferation or apoptosis elicited by the acting cytokine).

Per-sentence relation extraction process included several steps. First, we found verbs in sentences by looking for words tagged as VB, VBD, VBG, VBN, VBP or VBZ by part-of-speech tagger. For each verb, we examined all typed dependencies it was governing and attempted to resolve verb tense. In particular, we marked verbs that were tagged as VBN and that governed ‘passive nominal subject’, ‘passive auxiliary’ or ‘agent’ dependencies as passive. Second, to allow further relation-directionality detection, we performed cell–cytokine semantic linking by examining all possible entity combinations with a verb located between a cell and a cytokine. We considered a candidate (verb, cell and cytokine) tuple as semantically related, if the sentence contained a directional dependency tree path from one of the elements to the other two. For example, in the sentence “These results suggest that IL6 promotes IL22 secreting T_H17 subset differentiation” in Figure 1a, there is a path from the verb “promotes” to the cytokine IL-6 and to the T_H17 cell, allowing for (promote, IL-6, T_H17) relation identification. Third, we used the verb tense and the cell and cytokine order in the sentence to resolve relation directionality. In the example above, because verb tense is active, and the cytokine precedes the cell entity, we identified the cytokine as acting on the cell. Finally, we applied a manually assembled verb classification lexicon (Supplementary Table 5) to assign relation polarity (i.e., positive in the example above). Verb classification lexicon lookup was performed by using stemmed forms for both lexicon terms and the relation verb.

For the special case of cell–cytokine relations included within cell descriptions, ‘noun phrase–internal relations’ (e.g., ‘IFN- γ -producing CD4⁺ T cells’, ‘IL-2-activated NK cells’ or ‘NK cell IFN- γ ’), we applied a tailored rule-based approach (Supplementary Note 4).

If relation directionality could be deduced directly from the verb (e.g., express or stimulate), as marked by human curators of the verb lexicon, then we preferred the directionality denoted by the lexicon to directionality-related decisions made by the algorithm described above.

Finally, we applied cell entity-containing noun phrase analysis to detect cellular biological functions elicited by the interacting cytokine, such as ‘migration’ in the sentence “GM-CSF enhanced reactive oxygen species release and neutrophil migration *in vitro*”. We examined noun phrase words on the right of the recognized cell entity to look for one or more matches from a manually curated cellular function list.

In terms of terminology, we referred to each detected (cell entity occurrence, cytokine entity occurrence, directionality, polarity and cellular function) relation in a sentence as a ‘relation evidence record’, whereas summarization of evidence records through representative labels and IDs of the entities produced

unique candidate (cell, cytokine and directionality) interactions. Polarity and cellular function were ignored during summarization. We defined interaction ‘context’ by the disease entities that co-occurred in the same abstracts with the identified relation evidence records.

Filtering and scoring. We chose to put strong emphasis on precision over recall to ensure benefit and increase adoption by the community. To do so, we developed a confidence-scoring methodology for both individual relation evidence records and summarized interactions, and we performed four filtering stages as follows:

(i) Baseline filtering. For each putative relation evidence record (i.e., a directional cell–cytokine relation extracted from an individual sentence), we compiled a set of sentence-level features to capture the complexity of the sentence from which this evidence emerged. These included sentence length, number of typed dependencies, entities and relations detected, negation presence, cell ontology mapping score, as well as the distance between the cell and the cytokine occurrences. Next, we filtered evidence with potentially lower confidence, such as negated sentences or sentences with more than one non-noun phrase–internal relation detected or evidence with low cell ontology mapping score (**Supplementary Note 3**). This yielded a ‘baseline’ subset of putative interaction evidence records that was used in the subsequent scoring and filtering stages.

(ii) Evidence record confidence scoring. All individual records were assigned confidence scores, based on sentence-level features (**Supplementary Note 5**) to allow focus on the highest-confidence evidence, both when queried by iX web-based interface users and during analyses, as well as to serve as a feature for further filtering, as detailed below.

(iii) Summary scoring. All summarized interactions (i.e., unique cell–cytokine–directionality triples summarized across the corpus) were assigned with ‘summary’ and ‘enrichment’ scores (**Supplementary Note 6**) to allow focusing on highly cited interactions.

(iv) Lasso-based filtering of summarized interactions. Lastly, to choose optimal parameters for classification and further filtering of summarized interactions (i.e., unique cell–cytokine–directionality triples), we built a lasso logistic regression model⁵¹. Here we summarized evidence-level features into the interaction level across all evidence records, such as minimal or maximal sentence length, minimal or maximal cell–cytokine entity distance in sentence, the presence of mouse or human MESH term annotation in any of the evidence papers, and minimal or maximal evidence confidence scores described in (ii) above. Additionally, we defined interaction-level features, such as the overall number of evidence records, and the summary and enrichment scores defined in (iii) above. For all of the verbs in the training set, we added an interaction-level feature that reflected verb presence in any of the evidence records, which produced 178 features in total. We trained the lasso regression model on a set of 203 randomly selected interactions, summarizing 788 baseline evidence records identified by iX. We labeled the summarized interaction as positive if an only if at least one of its evidence records was manually classified as having cell, cytokine and directionality identified correctly. The resulting set of manually labeled (cell–cytokine–directionality) triples was used for training the lasso model, separately for incoming and outgoing interactions, with leave-one-out cross-validation. This procedure identified a linear combination of sparse feature weightings, which we then applied to all putative summarized interactions to classify them as either ‘true’ or ‘false’. The most prominent features included maximal evidence confidence score, having a ‘mouse’ MESH assigned for evidence record articles, as well as the presence of several verbs, such as ‘produce’, ‘synthesize’ and ‘affect’. We enforced classification of ‘true’ for interactions that had at least one noun phrase–internal evidence record, due to their very high identification precision. Filtering out summarized interactions that were classified as ‘false’, together with all of their evidence records, yielded the resulting data set, which we used for further performance evaluation and system-wide analyses (see **Supplementary Table 11** for a breakdown of the articles, records, entities and interactions remaining at various stages and **Supplementary Table 12** for PubMed articles for the resulting cell–cytokine relation evidence).

Named entity recognition (NER) performance evaluation. NER precision for all entity types was assessed by human curators. In addition, for cells, both

precision and recall were examined by comparing to the gold-standard Colorado richly annotated full text (CRAFT) corpus⁹ (**Supplementary Note 7**).

Relation precision evaluation. Relation extraction precision assessment was based on manual evaluation of randomly chosen relation evidence records (**Supplementary Note 8**).

Reference book network assembly and relation recall evaluation. To the best of our knowledge, no gold standard for directional cell–cytokine relations exists as of now. Thus, to assess iX text-mining process performance in capturing existing knowledge in general, and to evaluate relation recall in particular, 11 human curators manually annotated interaction mentions, specifying the cell and cytokine terms and interaction direction from a reference book with encyclopedic display of up-to-date knowledge about cytokines and their interactions²⁹. We mapped these to the Cell Ontology and cytokine lexicon, respectively, to summarize and acquire reference IDs consistent with those used in iX. This process yielded 725 unique (cell, cytokine and directionality) interaction triples, which we compared to those captured by iX, while allowing non-exact cell type match along the Cell Ontology hierarchy to account for the varying level of granularity of cell type reporting in the literature. This comparison assessed the proportion of reference book triples captured by iX, as well as the number of triples unique to either the reference book or iX. For the latter, to estimate false-positive proportion, we manually evaluated 200 (cell, cytokine, directionality) triples that were randomly selected from the iX knowledgebase and not covered by the reference book. An interaction was classified as false positive by the human curators, if none of the evidence records collected by iX for that triple reported the directional interaction in question. Following the assessment, we unified the list of interactions identified via literature text-mining and reference book annotation into a single knowledgebase.

Evaluation of co-occurrence-based relation extraction performance. A co-occurrence-based approach would not be able to capture interaction directionality, an essential characteristic of intercellular communication, as our typed dependency-based method inherently does. Still, we aimed to assess the quality of those easier-to-extract relations. To the best of our knowledge, no gold standard for cell–cytokine relations exists as of now. Thus, to evaluate co-occurrence-based relation extraction performance and contrast it with the typed dependency-based approach, we used the interaction network we manually assembled from the reference book (see the subsection “Reference book network assembly and relation recall evaluation”) and discarded interaction directionality for both reference-derived and typed dependency-based interactions (post-lasso filtering, hereafter referred to as ‘iX interactions’) to produce two sets of cell–cytokine pairs to compare with. We linked cell and cytokine entities, which we previously recognized as co-appearing within sentence boundaries, into relation evidence records and summarized them into interactions, with the number of distinct articles mentioning the relation defining the strength of evidence.

Cytokine degrees and power-law fit. The literature-derived nature of the iX network inherently reflects the fact that study of cellular interactions is performed at a varying level of granularity, rather than necessarily using the most specific cell type. As such, to avoid situations whereby the same interaction is counted multiple times when calculating degree distributions, we discarded interactions of less specific cell types, if a more specific cell type was known to interact with the same cytokine in the same direction. Cytokine degree counts, calculated separately per direction, took into account interactions with both hematopoietic and non-hematopoietic cells. Discrete power-law and log-normal distribution fit calculations were performed using the poweRlaw R package⁵².

Proteome comparison. We examined the similarity of iX and proteome interactions by inspecting each interaction direction separately and comparing cytokine degrees in these data sets. For outgoing interactions, we compared the number of distinct interacting cell types captured by iX and the number of those reported to express the particular cytokine in the ImmProt compendium³⁴. For incoming interactions, we examined the proteomic profiles of

genes encoding cytokine receptors and approximated cytokine–cell interactions by mapping expressed receptors to the respective cytokines, based on the KEGG “Cytokine–cytokine receptor interaction—*Homo sapiens* (human)” (hsa04060) pathway entry^{53,54}. We then compared the resulting proteome-derived degrees to those captured by iX for each cytokine. To avoid situations in which the same interaction was counted multiple times due to literature reporting at varying levels of cell-type granularity, we discarded iX interactions of less-specific cell types if a more specific cell type was known to interact with the same cytokine in the same direction. Moreover, to focus on similar cellular compartments, we discarded non-hematopoietic cell interactions from iX, thereby calculating cytokine degrees across hematopoietic cell types only in both data sets.

Cell–cytokine interaction prediction. To systematically predict interactions between immune cells and cytokines, we applied several strategies, separately for each interaction direction:

(i) Based on the similarity of global signaling profiles, we assembled a global literature-derived Boolean matrix that indicated whether an interaction has been described for each cell–cytokine pair. We used hierarchical clustering to group together cytokines with similar interaction profiles across all cell types and hypothesized that non-interacting cell–cytokine pairs located within highly connected clusters might actually interact. Therefore, we scanned clusters, confined to all possible combinations of column and row dendrogram branches, and derived interaction candidates for ‘gaps’ in clusters with at least 85% of interacting pairs. We scored the resulting candidates by counting the number of clusters that predicted the interaction to be novel.

(ii) On the basis of similarity of cytokine structure or function we calculated the proportion of cytokine family members known from the literature to interact with each cell type and derived interaction candidates by hypothesizing that cells interacting with at least 30% of the members of a cytokine family and with its most interactive member (i.e., usually the founding family member) might interact with other cytokines in the family as well. We used the aforementioned proportion to rank the resulting cell–cytokine interaction candidates.

(iii) On the basis of differences between literature-derived knowledge and experimental data (**Supplementary Note 9**).

Manual evaluation of cell–cytokine interaction candidates. We evaluated 78 interactions with the best overall scoring (**Supplementary Tables 14 and 15**). Candidates were scored by manual evaluation (e.g., we considered predictions made by multiple methodologies stronger), however, they incorporated multiple other criteria in defining which candidate interactions to validate. For example, we considered a candidate interaction whose reverse directionality was not reported yet to be stronger than the one for which it had been captured by iX (as that reverse directionality might result from erroneous identification, invalidating prediction novelty). A subset of candidate interactions with no evidence identified using a manual search was then chosen for experimental validation.

Experimental validation of cell–cytokine interaction predictions. Whole blood was obtained by consent from healthy volunteers through venipuncture. The PBMC fraction was separated by a standard centrifugation at 1,500 r.p.m. on a Ficoll gradient without applying the brake. For phospho-flow, cells were washed twice with Dulbecco’s PBS and subjected to stimulation for 15 min with IL-34 and CSF1 (Peprotech, Asia) at 100 ng/ml. Cells were fixed for 10 min at room temperature with 1.6% paraformaldehyde (PFA; Pierce) and stained for 1 h at room temperature with a mix of metal-tagged antibodies. Furthermore, cells were permeabilized with ice-cold methanol and subjected to intracellular staining with metal-tagged antibodies against a phosphorylated form of ERK1–ERK2, pNF-κB, pSTAT1 and pSTAT5. Antibodies to the phosphorylated targets were obtained from Fluidigm.

For intracellular cytokine staining, PBMCs were stimulated with PMA (150 ng/ml) and ionomycin (1 mM) (Sigma) for 4 h at 37 °C in complete medium containing monensin and brefeldin-A at a 1:1,000 dilution (eBioscience). Extracellular epitopes were stained for 1 h with a mix of metal-tagged antibodies, and cells were fixed with PFA as described above and permeabilized for 1 h on ice with saponin permeabilization buffer (eBioscience). Intracellular staining of IL-7 was performed on ice in saponin-containing

buffer. All extracellular and cytokine-specific antibodies were conjugated in-house using MaxPar kits from Fluidigm or pre-conjugates purchased from Fluidigm. Cells were stained with Cell-ID Ir191/193 for viability stain and acquired on a CyTOF1 (DVS, Fluidigm) instrument.

Assembly of disease similarity modules. We defined a context for cells, cytokines and cell–cytokine interactions by diseases co-occurring in the same abstracts, whereas disease mentions were captured by using a manually curated compilation of the UMLS Metathesaurus⁴⁹. To identify disease-similarity modules, we focused on the 188 top-cited diseases (co-mentioned in at least 500 papers with cytokines and at least 500 papers with hematopoietic cells) that we could classify as pertaining to at least one of the eight predefined clinical categories (e.g., disorder of cardiovascular system, generalized metabolic disorder, neoplasm of hematopoietic and nonhematopoietic cell types, autoimmune diseases and hypersensitivity conditions). To define clinical categories, we used SNOMED CT ontology (available through UMLS Metathesaurus⁴⁹) and manually expanded its autoimmune disease category with a publicly available autoimmune-related disease list (<http://www.aarda.org/autoimmune-information/list-of-diseases>). As a preliminary step to module assembly, we extracted a nonspecific across-disease control profile by repeated paper sampling from the entire corpus of 521,625 disease–HPC and 438,012 disease–cytokine co-occurrence papers, respectively, without limiting to any particular context (200 iterations of 200 papers each). We examined cells and cytokines mentioned in the sampled papers, assembling the distribution of hits for each HPC and cytokine across sampling iterations, to serve a control for further disease-specific profile assembly.

Next, for each of the 188 diseases of interest, we assembled its underlying signaling profile by applying several steps: (i) performing 200 samplings of 200 random papers from the disease-specific sub-corpus to control for differences in corpus size, followed by extraction of the distribution of hits for each cell and cytokine across sampling attempts, (ii) filtering out under-represented entities, i.e., those with a hit proportion median lower in the disease-specific sampling than in the control sampling and (iii) linking cells and cytokines that co-occur in the same disease profile to interacting pairs, using interaction potential captured in the overall iX network.

Lastly, we performed unsupervised clustering of the resulting Boolean disease profiles (‘0’ or ‘1’ to indicate whether the particular cell, cytokine or pair was a part of the profile) using the WGCNA R package^{55,56} to assemble the set of immune-centric disease similarity modules based on the binary distance metric.

Disease module signature extraction. To extract features (i.e., cells, cytokines and their interacting pairs) that characterized disease modules, we applied a hypergeometric test, independently for each feature, examining whether the feature was over-represented in the particular module as compared to that in the entire set of 188 diseases. We corrected for multiple testing using a Benjamini–Hochberg correction.

Cytokine–disease association prediction. To test cytokine utilization across conditions, we repeatedly sampled papers from disease-specific sub-corpora, separately for each of the 188 top-cited diseases (200 iterations of 200 papers each). We extracted the distribution of paper hit proportions for each cytokine–disease pair across iterations and used the median proportion as the measure of cytokine mention frequency in that particular context.

To systematically predict cytokine–disease associations, we employed global within-module immune similarity and applied the following steps separately for each disease module: (i) assembly of literature-derived Boolean matrix indicating whether a cytokine was a part of the disease profile, for each disease in the module and (ii) hierarchical clustering to group together module diseases displaying similar profiles across all cytokines and hypothesizing that an association should exist between currently unlinked cytokine–disease pairs within highly linked clusters. Therefore, we scanned clusters, which were confined to all possible combinations of column and row dendrogram branches, and derived cytokine–disease association candidates for ‘gaps’ in clusters with at least 30% of linked cytokine–disease pairs. We scored the resulting candidates by counting the number of clusters that predicted the particular association to be novel.

Life Sciences Reporting Summary. Further information on experimental design is available in the **Life Sciences Reporting Summary**.

Data availability. Data of the intercellular communication network and disease context are hosted on ImmPort with periodic updates and are available for query and download with standardization to ontologies at <http://www.immunexpresso.org>.

46. Hughes, A.L. *Vertebrate Immune System: Evolution* (John Wiley and Sons, Ltd., 2001) <http://dx.doi.org/10.1002/9780470015902.a0006125.pub2>.
47. Du Pasquier, L. The immune system of invertebrates and vertebrates. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **129**, 1–15 (2001).
48. De Marneffe, M.-C., MacCartney, B. & Manning, C.D. Generating typed dependency parses from phrase structure parses. *LREC Int. Conf. Lang. Resour. Eval.* **6**, 449–454 (2006).
49. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267–D270 (2004).
50. Bard, J., Rhee, S.Y. & Ashburner, M. An ontology for cell types. *Genome Biol.* **6**, R21 (2005).
51. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1994).
52. Gillespie, C.S. Fitting heavy-tailed distributions: the ‘poweRlaw’ package. *J. Stat. Softw.* **64**, 1–16 (2015).
53. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
54. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
55. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
56. Langfelder, P. & Horvath, S. Fast R functions for robust correlations and hierarchical clustering. *J. Stat. Softw.* **46**, 1–17 (2012).

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see [Reporting Life Sciences Research](#). For further information on Nature Research policies, including our [data availability policy](#), see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study. For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

► Experimental design

1. Sample size

Describe how sample size was determined.

1. For validation of cytokine-disease association prediction (Fig. 6b, 6c), sample size was dictated by the respective datasets published on GEO. Data exclusion to ensure pairing produced lists of 24 and 28 independent samples, which should be sufficient for two-sided Wilcoxon signed rank test performed.
2. For experimental cell-cytokine interaction validation we obtained blood from 6 healthy donors and results were reproduced sufficiently.

2. Data exclusions

Describe any data exclusions.

1. For cell-cytokine interaction prediction validation (Fig. 4c, 4d)
2. For validation of cytokine-disease association predictions (Fig. 6b, 6c), since paired statistical test is performed, only psoriatic patients having data for both lesion and healthy skin at baseline are included. The rest of the samples available in the public gene expression datasets examined for this experiment, are excluded.

3. Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

1. IL7 production by monocyte, dendritic cell and T cell subsets (Fig. 4c) was repeated on 2 healthy donors, reproducing the experimental findings
2. pNFKB and pSTAT5 phosphorylation in T cell subsets, following IL34 stimulation (Fig. 4d) was repeated on 5 healthy donors, reproducing the experimental findings as well.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Irrelevant to the study, as cell-cytokine interaction prediction validation involved no allocation into experimental groups

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Irrelevant to the cell-cytokine interaction prediction validation experiment, as no group allocation was performed

Note: all *in vivo* studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a Confirmed

- The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
- A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- A statement indicating how many times each experiment was replicated
- The statistical test(s) used and whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of any assumptions or corrections, such as an adjustment for multiple comparisons
- Test values indicating whether an effect is present
Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.
- A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
- Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on [statistics for biologists](#) for further resources and guidance.

► Software

Policy information about [availability of computer code](#)

7. Software

Describe the software used to analyze the data in this study.

Python 2.7, Stanford Parser v2.0.4, ElasticSearch v2.3.5 and MySQL 5.7.17 were used for building the knowledgebase. UMLS Metathesaurus v2013AA and Cell Ontology release 2013-04-20 served a basis for disease and cell ontologies. R software environment (R Studio v1.0.44, R v3.3.2) was used for statistical computing and graphics, with packages further specified in the Online Method sections. For experimental data analysis we used Cytobank

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

► Materials and reagents

Policy information about [availability of materials](#)

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.

Metal-tagged Ab's for mass cytometry were obtained as pre-conjugates from Fluidigm or purchased in a purified form from other vendors and tagged using Fluidigm X8 and DN3 conjugation kits. All antibody clones are available for purchase from stated vendors. In-house conjugates are available in limited quantity at the lab.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

CD8a	Nd146	Fluidigm
CD3	Er170	Fluidigm
CCR7	Tb159	Fluidigm
CD45RA	Dy161	Biolegend Clone HI100
NF-KB	Er166	Fluidigm
pSTAT1	Eu153	Fluidigm
pSTAT3	Gd158	Fluidigm
pSTAT5	Nd150	Fluidigm
pCREB	Ho165	Fluidigm
CD69	Dy162	Biolegend Clone FN50
CD25	Tm169	Fluidigm
PERK	Er177	Fluidigm
IL7	Yb171	
CD11b	Yb172	
CD14	Gd160	

Antibodies were purchased from Fluidigm or conjugated and titrated in-house on healthy donor PBMC samples according to Fluidigm protocols.

10. Eukaryotic cell lines

- a. State the source of each eukaryotic cell line used.
- b. Describe the method of cell line authentication used.
- c. Report whether the cell lines were tested for mycoplasma contamination.
- d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No cell lines used

No cell lines used

No cell lines used

No cell lines used

► Animals and human research participants

Policy information about [studies involving animals](#); when reporting animal research, follow the [ARRIVE guidelines](#)

11. Description of research animals

- Provide all relevant details on animals and/or animal-derived materials used in the study.

No animal studies were carried out

Policy information about [studies involving human research participants](#)

12. Description of human research participants

- Describe the covariate-relevant population characteristics of the human research participants.

Healthy donor whole blood was obtained by venipuncture according to institutional IRB. Blood was collected from 6 healthy donors in total. 4 males and 2 females age 35-45yo.