Pacific Symposium on Biocomputing 14:439-450 (2009)

TOWARDS A CYTOKINE-CELL INTERACTION KNOWLEDGEBASE OF THE ADAPTIVE IMMUNE SYSTEM

SHAI S. SHEN-ORR[1,2], OFIR GOLDBERGER[2], YAEL GARTEN[3], YAEL ROSENBERG-HASSON[4], PATRICIA A. LOVELACE[4,5], DAVID L. HIRSCHBERG[4], RUSS B. ALTMAN[6], MARK M. DAVIS[2,7], ATUL J. BUTTE[1,8]

[1] Stanford Biomedical Informatics Research, Department of Medicine, [2] Department of Microbiology & Immunology,[3] Stanford Biomedical Informatics Training Program,[4] Stanford Human Immune Monitoring Center, [5] Stem Cell Biology and Regenerative Medicine,[6] Departments of Bioengineering and Genetics,[7] The Howard Hughes Medical Institute,[8] Department of Pediatrics
Stanford University, Stanford, CA 94305-5479, USA

The immune system of higher organisms is, by any standard, complex. To date, using reductionist techniques, immunologists have elucidated many of the basic principles of how the immune system functions, yet our understanding is still far from complete. In an era of high throughput measurements, it is already clear that the scientific knowledge we have accumulated has itself grown larger than our ability to cope with it, and thus it is increasingly important to develop bioinformatics tools with which to navigate the complexity of the information that is available to us. Here, we describe ImmuneXpresso, an information extraction system, tailored for parsing the primary literature of immunology and relating it to experimental data. The immune system is very much dependent on the interactions of various white blood cells with each other, either in synaptic contacts, at a distance using cytokines or chemokines, or both. Therefore, as a first approximation, we used ImmuneXpresso to create a literature derived network of interactions between cells and cytokines. Integration of cell-specific gene expression data facilitates cross-validation of cytokine mediated cell-cell interactions and suggests novel interactions. We evaluate the performance of our automatically generated multi-scale model against existing manually curated data, and show how this system can be used to guide experimentalists in interpreting multi-scale, experimental data. Our methodology is scalable and can be generalized to other systems.

1. Introduction

Modern biology has benefited greatly from reductionism, perhaps most notably in the last century by the decision by Delbruck and colleagues to focus on the lowly bacteriophage in their efforts to discern the relationship between genes and DNA on a chromosome, thus escaping the complexities of higher organisms. But to understand the complexities of systems that are only present in higher organisms, such as the immune system, reductionism can only be taken so far before it veers into oversimplification. This is especially true in human immunology, where it is much more difficult to perform experiments as tightly controlled with respect to the many possible variables as in the mouse model. For these reasons we believe that high throughput assays and their accompanying large datasets will be increasingly the norm in immunology, and thus it is critical to develop bioinformatics tools that can assimilate and interpret these datasets as efficiently and broadly as possible, ideally not just the output of one laboratory, but of thousands, and over the several decades that could be considered the modern era.

With respect to how this might be accomplished, the complex and concise nature of the scientific literature means that the use of information extraction tools developed for generic tests is often insufficient. Driven by the utility of high-throughput technologies such as microarrays to measure whole genome gene expression data and yeast two hybrid screens to measure protein-protein interactions, many of the information extraction systems were developed with the aim of extracting genetic[1] or protein-protein interactions[2] from the literature or for annotating groups of differentially expressed genes. Results from these have been used either for construction of searchable knowledgebases[1,3] or for validation of high-throughput experimental results[4].

Here, we construct an information extraction system, partly based on the earlier Textpresso[1] and later Pharmspresso[3] knowledgebases, which we have named ImmuneXpresso, to search abstracts of primary immunology literature for interactions between adaptive immune cells and the cytokines which they secrete and/or affect them. Using the identified interactions, we assemble an inter-cellular network of cells and cytokines to which we integrate with cell type specific gene expression data. Identification of expressed cytokines and receptors in specific cells provides support for ImmuneXpresso identified interactions and allows, in many instances, to assign them directionality. We evaluate the performance of our automatically generated model against existing manually curated data, and use it to guide the identification of novel findings.

2. Methods

2.1 *Immune related corpus and lexicon for information extraction*

To define a comprehensive list of relationships between cells and cytokines, we queried the NCBI journals database to identify journals reporting findings in immunology. 326 journals are annotated under the subject terms: “Immunology & Allergy”, or “General Science”. Our corpus thus consists of all abstracts published in these journals circa-1960 (PubMed limitation) and onwards, downloaded using NCBI's *eFetch* utility. We established a lexicon of immunological terms, and a comprehensive list of their synonyms (e.g., RANTES = CCL5). Our lexicon is currently limited to six adaptive immune system cell types: B-cells, Cytotoxic T-cell lymphocytes (CTLs), T helper cells, T-regulatory cells, γδ T-cells, and dendritic cells) and 38 cytokines and growth factors. We also constructed a list of verb stems which describe interactions between terms (e.g., induc, stimulat, repres).

2.2 *Automatic construction of immune-centric knowledgebase of cells and cytokines*

To extract information from the primary literature on interactions of immune system components, we built on the Textpresso extracting and processing package[1]. Given a corpus of literature and a lexicon of concepts, ImmuneXpresso identifies occurrences and co-occurrences of concepts within those sentences as well as relations between them. Importantly, owing to its rules, lexicon and corpus, ImmuneXpresso is tailored for the immunology knowledge domain with its unique and developing jargon and expressions. Our corpus of immune related text was tokenized into single sentences. For each sentence we marked all appearances of terms from our lexicon of cells, cytokines and verbs. We then identified all sentences in which two or more terms of different categories co-occurred (e.g. IL-2 stimulates T-helper cells). Next, a filtering step was applied to remove sentences not containing interaction terms or sentences from which it was difficult to infer a regulatory interaction, such as those containing negation. We categorized the verbs in our lexicon as describing positive or negative interactions. For example, sentences containing interactions described using terms such as 'stimulate', 'increase' and 'induce' were all considered descriptive of a positive interaction, whereas sentences containing terms such as 'decrease', 'interferes with production of' or 'deactivates' were categorized as describing negative interactions. A third category of 'undetermined' interactions were those for which we could not assign either a positive or negative role. These included sentences in which terms such as 'alter', 'mediates' and 'cooperatively' appeared. The resulting output is a list of all interactions ImmuneXpresso detected in the corpus. Each interaction described a relationship of a certain type (positive, negative or undetermined) between a cell and a cytokine. For each we also kept track of the number of sentences an interaction was detected in, for use as measure of confidence in an interaction. The resultant interactions were visualized as a bipartite network of cells and cytokines using Cytoscape[5].

2.3 *Data for cell specific gene expression*

Ultimately, cellular communication through cytokines is mediated by the binding of a cytokine secreted from one cell to a receptor expressed on the surface of another. The signature of cytokine and receptor expression should be detectable in cell specific gene expression patterns. We identified publically available microarray studies in GEO in which specific cell subsets were isolated from human blood and their gene expression measured. Data for stimulated and unstimulated dendritic cells was obtained from Jeffery et al.[6] (GSM IDs: 90666, 90836, 90837, 90664). Data for T-regulatory cells was obtained from Ocklenburg et al.[7] (GSM IDs: 101521, 101519). In both of these studies, hybridizations were done on an Affymetrix GeneChip Human Genome U133 Array Set HG-U133A. Palmer et al. [8] (GSE4889) isolated B-cells, T-helper cells and CTLs from healthy individuals and hybridized them on an array. These are two-color printed microarrays in which a single individual's cells (Cy5 labeled) were hybridized against a standard reference – a mixture of RNA from 11 human cell lines. Blood samples from 3 males and 3 females were drawn, and cells selected for using magnetic beads. B-cell selection was positive, whereas for T-helper and CTLs selection was negative. To the best of our knowledge, no high quality γδ T-cell study for humans is publicly available.

For each cell type, we performed quantile normalization[9] across all replicate and/or multiple measurements. We matched microarray platform probes to Entrez GeneIDs[10]. Those probes not mapped to any Entrez GeneID were discarded, whereas the intensity values of those mapping to the same gene were averaged. We converted the expression values into binary form by considering, for each array, only the genes in the top 40% of the intensity values, as expressed. We picked the 40th percentile as this is where the log ratio turns positive on most two color arrays in this data. For each cell type, we generated a cell type specific gene expression signature by considering a gene as expressed if it was detected as expressed in arbitrarily 50% or more of the replicate or multiple array measurements of that cell type. This resulted in 4338, 4468, 4931, 5803 and 6122 genes considered as expressed in B-cells, T-helper cells, CTLs, dendritic cells and T-regulatory cells, respectively.

2.4 *Linking inter-cellular interactions with intra-cellular content*

To link our inter-cellular communication network to the intra-cellular networks which drive them, for each cytokine listed in the ImmuneXpresso lexicon, we identified in the Entrez Gene database one or more receptors to which it binds.

This complete list included a total of 79 genes, 38 of which were cytokines or growth factors and 41 receptors, with 68 binary cytokine-receptor interactions between them. For example, the binding of IL-13 to the IL13RA1 receptor as well as IL13RA2 and IL4RA was considered as three separate binding events. All 79 genes had probes on at least one of the array platforms (see 2.3), but only 57 had probes on both array platforms. We use these to compile a cellular gene expression driven cell-cytokine network and contrast it with a literature-derived network generated by ImmuneXpresso.

2.5 Validation and Experimental data

It is critical to assess the quality of the output ImmuneXpresso extracts, namely how well do we automatically extract cell-cytokine interactions from literature. The Cytokine Online Pathfinder Encyclopedia (COPE) database [11] and the Cytokine Reference – Online Database [12] are the two largest publicly available references for cytokine functionality. Both are manually curated by experts. We used these databases to search for each of the cell-cytokine interactions identified by ImmuneXpresso and check its validity.

Unlike the free text which it mines, ImmuneXpresso output is in machine interpretable form that can easily be utilized for downstream analysis. We contrasted ImmuneXpresso output to two experimentally derived datasets: in the first, blood was drawn from 29 human individuals, males and females of varying ages, and analyzed for cell frequency and serum cytokines. The frequency of B-cells, γδ T-cells, CTLs and T-helper cells was assessed using antibody staining and flow cytometry. Serum cytokines were analyzed using the Luminex 200 assay system with MilliPore 37-plex kits. Employing our network algorithm (relevance networks)[13] and using a Pearson's correlation cutoff of ±0.8, we identified 668 associations in the data, of which 41 were between cells and cytokines.

In the second dataset, a cytokine response assay was performed on CD4+ T cells from spleens of 10 week old MRL.MpJ mice. These were stimulated with various dilutions of 5 cytokines (IL-2, IL-3, IL-12, IFN-γ and TNF-α) by incubation for 16 hours at 10^6 cells per well. Secretion blockers were then added for an additional 2 hours in order to accumulate intracellular cytokines and then the cells were lysed and lysates used for Luminex cytokine multiplex assays with Panomics Procarta 21-plex kits. An increase, appearance or decrease in the abundance of the 21 measured cytokines compared with their abundance with no stimulation, was indicative of an increased production above background, *de novo* production and inhibition of production of cytokines respectively.

3.Results

A total of 3695 sentences reporting interactions between the six cell types used here and 38 cytokines were identified by ImmuneXpresso (Fig. 1). These were unified into 217 unique interactions, with an average number of 17 sentences per interaction. The resultant network is dense, with 54% of cell-cytokine pairs connected to one another directly. In comparison with other biological networks, this is a very dense network[14], in agreement with the high density value observed for the cellular immune network (nodes are cells, edges are cytokines) that Frankenstein et al.[14] derived from curated databases.

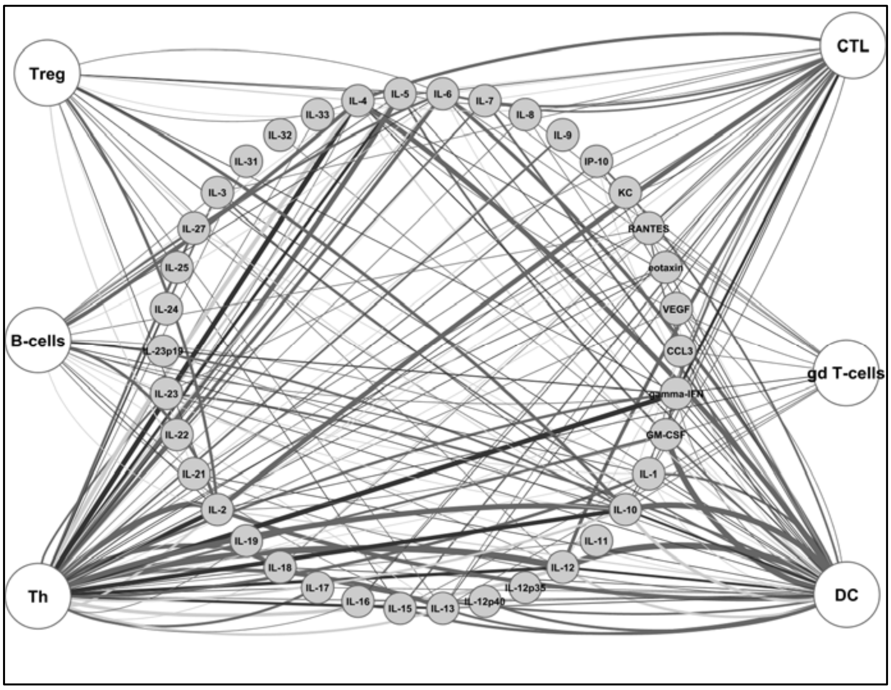

Figure 1.An automatically derived network of cells and cytokines from the immunological literature. Six immune system cell subsets (white nodes) and their relationship, through 217 edges, with 38 cytokines (grey nodes). Edge color, from light to dark denotes the type of interaction, positive, undetermined, and negative. Edge width denotes the confidence of the relationship, represented as the log of the number of sentences in which the interactions was identified. Treg – T-regulatory cell, Th – T-helper cell, γδ – gamma delta, DC- dendritic cell.

Despite the high density we observe, the theoretical density which could be reached with 217 edges in a bipartite network of this size is 95%. However, many of the cell-cytokine pairs have more than one edge type spanning between them which likely reflects either incorrect semantic parsing of the text or the multiple functionality of cytokines under different conditions or in specific cell subsets. From the 217 edges, 113 were positive, 50 were negative and 54 were of undetermined type. Examination of the number of sentences shows that the evidence for most cell-cytokine relationships is predominantly of one type (Fig. 2). Furthermore, there is a strong correlation (Pearson's 0.91 for cell types, 0.59 for cytokines) between the degree of a node and the number of sentences which support its interactions. This is to be expected in a literature based network as the more research is done in a field, the more separate components become associated with one another.

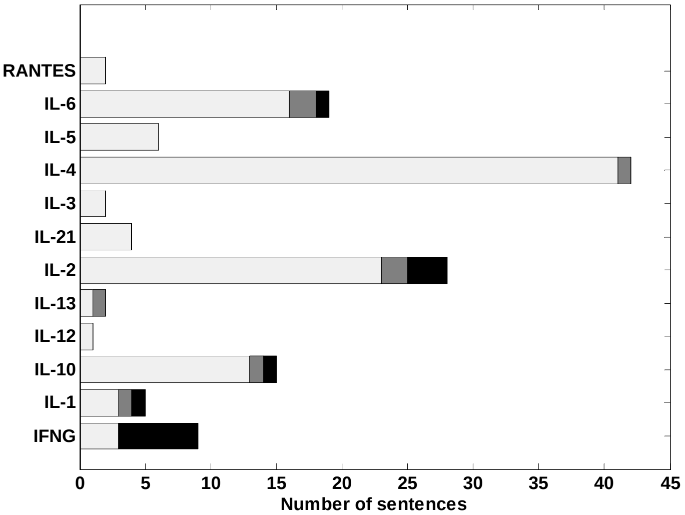

Figure 2: Interactions between B-cells and twelve cytokines extracted from abstracts by ImmuneXpresso. X-axis shows the number of sentences the interaction was observed in. Color denotes interaction type. White, gray and black denote positive, negative and undetermined interaction respectively. All twelve cytokines are true effectors or products of B-cells.

To evaluate our ability to correctly capture cell-cytokine interactions automatically from abstracts, we compared our results to the relevant subset of a manually curated set of cytokine-cell interactions from the COPE database and the Online Cytokine Reference[12]. Checking specifically by cell type, we verified each of our detected B-cell interactions with those available on the two websites. Overall, we had a 40% false-negative rate, but no false positives, better results than are usually observed for co-occurrence systems which search abstracts only[1]. Two of the cytokines interacting with B-cells, IL-21 and RANTES, did not appear in the Online Cytokine Reference, but did appear in COPE. Similar false negative rates were observed for dendritic cells. Interestingly, more recent discoveries in cytokine-cell regulation such as the regulatory interactions of γδ-T-cells, or IL-33 did not appear in either database, showing the power of automatic annotation for keeping up to date with published literature.

Edges in the ImmuneXpresso network link cells and cytokines to one another (Fig. 1). In reality, these cells produce and secrete cytokines which bind to membrane receptors expressed on the surface of another cell or even the same one. We hypothesized that inter-cellular interactions between cytokines and their receptors would be evident, in cell specific gene expression data. If we could detect cytokines and receptors gene expression, than we may be able to assemble gene-expression-based inter-cellular communication networks. Further, if cytokines and their binding partner receptors were exclusively expressed between cells, we may be able to assign directionality to currently undirected ImmuneXpresso reported interactions. To test this, we assembled a compendium of cell specific gene expression signatures from publically available microarray studies (see 2.3) matching the cell types in the present version of the ImmuneXpresso lexicon, as well as identified the one or more receptors each cytokine binds (see 2.4).

We asked how many cytokines in a given cell type are exclusively expressed from their receptor and how many are expressed in the same cell. The majority of cytokines and receptors were expressed exclusively of their binding partner. For example, the 8 cytokines and 21 receptors we detect expressed in T-regulatory cells may participate in 35 different binary interactions, both within our model framework (5 immune cell subsets, no data for γδ-T-cells) and outside, but for only two (IL-13 binds IL4RA and IL-16 binds CD4) are both the receptor and the cytokine expressed by T-regulatory cells. We note that the cytokine IL-2, known to both be expressed by T-regulatory cells, and regulate them via IL2Ra (CD25), is not detected as expressed in T-regulatory cells under our criteria. Similarly, for only 7 out of 40, 10 out of 30, 6 out of 40 and 5 out of 34 for B-cells, T-helper, CTL and dendritic cells respectively, both cytokine and receptor are expressed on the same cell. Expression of both cytokine and receptor in the same cell may be indicative of auto-regulation, or of the expression of either the cytokine or its' receptor under different conditions or subsets.

The ImmuneXpresso network is bipartite, with each edge representing an interaction between cell and cytokine. Each edge in this network has been extracted from the literature and thus is supported by experimental evidence. On the other hand, an interaction between two cells must always consist of at least two edges. Unlike single edges, interpretation of a path of two or more edges as a cytokine mediated interaction between two cells, is not necessarily warranted, as it requires one cell to be a producer of the cytokine and the other to be affected by it. Here, gene expression data may come to the aid, as it allows one to identify cells expressing or producing cytokines, and those with the potential to be affected by them by expressing the corresponding cytokine receptors.

Ignoring edge types, the ImmuneXpresso network has 119 edges. In theory, these could encode for 309 possible undirected regulatory paths between the 6 cells in our system, 119 of which are auto-regulatory. Using the information obtained from the gene expression data, we can now estimate how many of the pathways theoretically derived from the ImmuneXpresso output are supported by the gene expression data, such that a cytokine and its receptor are expressed in the two communicating cells. To do so, we consolidated the gene specific cytokine-receptor information we obtained from Entrez to match the details of the lexicon ImmuneXpresso uses. For example, the information 'B-cells express both IL-10 and the IL-10 receptor IL-10Ra and IL-10Rb' is simplified in the consolidated form to 'B-cells express the IL-10 cytokine and a receptor to which it can bind'.

Filtering the 309 ImmuneXpresso paths by requiring one of the cells to express a cytokine while the other expresses its receptor drops the number of interactions to 158 of which 30 are auto-regulatory (same cell expresses both the cytokine and the receptor). The remaining 151 non-functional paths are either a byproduct of the network representation, or appear as such due to the threshold we set as to which genes should be considered expressed (see 2.2). Furthermore, as we observed, many of the cytokines and receptors are exclusively expressed on one of the two cells. Therefore, unlike the directionless interactions ImmuneXpresso currently reports, many of the cytokine mediated cell-cell interactions we infer from the gene expression data are directional. Of the 158 ImmuneXpresso paths supported by the gene expression data, we can assign directionality to 76. Last, we can ask the reverse question, namely how many of the possible cytokine mediated, cell-cell interactions appear in the gene expression data, and cannot be traced to any path in ImmuneXpresso. We find 27 such paths, 21 of which we could not find in the manually curated cytokine databases[11,12]. Each represents a hypothetical cytokine mediated cell-cell interaction that can be tested experimentally.

The utility of a knowledgebase in machine computable format stands out when conducting high throughput discovery driven experiments. In such experiments, researches rarely have expert knowledge in all of the variables being assayed and the number of results from experiments is often very high. Thus, it may be difficult to prioritize findings and link them to one another or to previous discoveries, to establish a comprehensive perspective. As a proof of principle, we analyzed serum cytokine and cell subset frequency data measured at the Stanford Human Immune Monitoring Core for 29 individuals, males and females of varying ages. Of the 41 detected interactions (see 2.5), 18 were between a cell and a cytokine covered in the present ImmuneXpresso lexicon version (see Data sources 2.1). Remarkably, ImmuneXpresso could verify each of those 18 and match it with a reported interaction in the literature. For example, an interaction between IL-15 and CTLs, which was deduced by our algorithm by the positive correlation observed between IL-15 and CTLs, was detected 9 times in abstract sentences by ImmuneXpresso, from such sentences as ‘These findings identify a novel CTL costimulatory pathway regulated by IL-15 and suggest that tissues can fine-tune the activation of effector T cells based on the presence or absence of stress and inflammation’[15]. Comparison of a second dataset, in which interactions of 13 cytokines were experimentally identified with spleen derived CD4+ T-helper cells, showed that ImmuneXpresso could validate 10 out of the 13 observed interactions. Manual searches for the other three interactions was able to confirm one additional interaction not captured by ImmuneXpresso. Each such interaction suggests a testable hypothesis which requires considerable time and resources to test. As datasets grow larger, machine identification and prioritization of novel findings to follow up on is key.

4. Discussion

The multi-scale nature of the immune system and its high complexity challenge us to find new ways to analyze it. Discoveries in immunology and cellular biology are occurring at an unprecedented rate. Despite this, we are far away from understanding how the immune system of higher organisms, and humans in particular, are able to mount an immune response. Here, we take the first steps towards building bioinformatics tools that are specifically geared towards extracting information from primary immunology literature and comparing it to high throughput data. For the primary literature in this field, a particular value of ImmuneXpresso is that it facilitates 'connecting the dots' between very different areas of immunology. Its low false positive rate, unlike the 35% false positive rate[1] observed in extraction systems specializing in genetic interactions, may be due to the types of associations we are extracting (i.e. cells with cytokines) which may be easier to extract from natural language text than other types of associations. The false negative rate matches that of other systems using abstracts to extract information. Expansion to full text analysis is likely to drop the false negative rate[1], but will likely result in an increase in false positives.

However, solely relying on information extraction systems to establish interaction networks always run the risk of misleading. As the accuracy of each individual reported interaction may not to be perfect, confidence in correctness of any given path decreases as a function of path length. To aid in this, as well as validate our findings, we use cell specific gene expression data to establish a biological context to our network. These provide both supporting evidence for extracted information and utilization of the extracted information for the identification of novel associations. We note that at present, our approach discards low abundance genes, and is sensitive to the conditions under which the gene expression data was assayed. These limit our ability to infer on cytokine mediated cell-cell interaction. Furthermore, extension of this approach to rare cell types may be complicated by the smaller body of literature and the lack of expression data. Thus we advocate systems integrating data from multiple data sources. For example, the requirements for high accuracy in natural language text processing, may be lifted to some extent by the integration of machine formatted data from other electronic sources. Future work should address this by integration of protein interaction data, enriching the lexicon by the use of predefined ontologies, and expanding the system to integrate not only cytokines and receptors, but also activation pathways and downstream targets.

Acknowledgments

We thank Alex Skrenchuk and Rong Chen from Stanford University for computer support and database support, respectively. The Ellison Foundation and the NIH specifically U19 AI057229-06, as well as the Hedco and the Russell Family Foundation for funding HIMC research. YG is supported by training grant NIH LM007033 from the National Library Medicine. RBA is supported by NIH/NIGMS Pharmacogenetics Research Network and Database and the PharmGKB resource (NIH GM61374). AJB is supported by the Lucile Packard Foundation for Children's Health.

Reference

1. Muller, H.M., Kenny, E.E. & Sternberg, P.W. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol* **2**, e309 (2004).
2. Rzhetsky, A. et al. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J Biomed Inform* **37**, 43-53 (2004).
3. Garten, Y. & Altman, R.B. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC Bioinformatics* (Accepted).
4. Rual, J.F. et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-8 (2005).
5. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-504 (2003).
6. Jeffrey, K.L. et al. Positive regulation of immune cell function and inflammatory responses by phosphatase PAC-1. *Nat Immunol* **7**, 274-83 (2006).
7. Ocklenburg, F. et al. UBD, a downstream element of FOXP3, allows the identification of LGALS3, a new marker of human regulatory T cells. *Lab Invest* **86**, 724-37 (2006).
8. Palmer, C., Diehn, M., Alizadeh, A.A. & Brown, P.O. Cell-type specific gene expression profiles of leukocytes in human peripheral blood. *BMC Genomics* **7**, 115 (2006).
9. Irizarry, R.A. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-64 (2003).
10. Chen, R. & Butte, A.J. AILUN: reannotating gene expression data automatically. *Nature Methods* **4**, 879 (2007).
11. Ibelgaufts, H. COPE: Cytokines Online Pathfinder Encyclopaedia. . (1997).
12. Oppenheim JJ et al. The Online Cytokine Reference Database. (2000).
13. Butte, A.J., Tamayo, P., Slonim, D., Golub, T.R. & Kohane, I.S. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci U S A* **97**, 12182-6 (2000).
14. Frankenstein, Z., Alon, U. & Cohen, I.R. The immune-body cytokine network defines a social architecture of cell interactions. *Biol Direct* **1**, 32 (2006).
15. Roberts, A.I. et al. NKG2D receptors induced by IL-15 costimulate CD28-negative effector CTL in the tissue microenvironment. *J Immunol* **167**, 5527-30 (2001).