Bioinformatics, 31(19), 2015, 3147–3155
doi: 10.1093/bioinformatics/btv316
Advance Access Publication Date: 27 May 2015
Original Paper

OXFORD

Gene expression

xHeinz: an algorithm for mining cross-species network modules under a flexible conservation model

Mohammed El-Kebir[1,2,3,†], Hayssam Soueidan[4,†], Thomas Hume[5,6,†], Daniela Beisser[7], Marcus Dittrich[8,9], Tobias Müller[8], Guillaume Blin[6], Jaap Heringa[2], Macha Nikolski[5,6], Lodewyk F. A. Wessels[4] and Gunnar W. Klau[1,2,10,*]

[1]Life Sciences, Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands, [2]Centre for Integrative Bioinformatics VU, VU University Amsterdam, The Netherlands, [3]Center for Computational Molecular Biology, Brown University, Providence, RI, USA, [4]Computational Cancer Biology, The Netherlands Cancer Institute, Amsterdam, The Netherlands, [5]Univ. Bordeaux, CBiB, 33000 Bordeaux, France, [6]Univ. Bordeaux, CNRS/LaBRI, 33405 Talence, France, [7]Genome Informatics, Institute of Human Genetics, University Hospital Essen, University of Duisburg-Essen, [8]Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg, Germany, [9]Institute of Human Genetics, University of Würzburg, Würzburg, Germany and [10]Erable Team, INRIA, Lyon, France

Associate Editor: Ziv Bar-Joseph

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Received on October 20, 2014; revised on May 12, 2015; accepted on May 18, 2015

Abstract

Motivation: Integrative network analysis methods provide robust interpretations of differential high-throughput molecular profile measurements. They are often used in a biomedical context—to generate novel hypotheses about the underlying cellular processes or to derive biomarkers for classification and subtyping. The underlying molecular profiles are frequently measured and validated on animal or cellular models. Therefore the results are not immediately transferable to human. In particular, this is also the case in a study of the recently discovered interleukin-17 producing helper T cells (Th17), which are fundamental for anti-microbial immunity but also known to contribute to autoimmune diseases.

Results: We propose a mathematical model for finding active subnetwork modules that are conserved between two species. These are sets of genes, one for each species, which (i) induce a connected subnetwork in a species-specific interaction network, (ii) show overall differential behavior and (iii) contain a large number of orthologous genes. We propose a flexible notion of conservation, which turns out to be crucial for the quality of the resulting modules in terms of biological interpretability. We propose an algorithm that finds provably optimal or near-optimal conserved active modules in our model. We apply our algorithm to understand the mechanisms underlying Th17 T cell differentiation in both mouse and human. As a main biological result, we find that the key regulation of Th17 differentiation is conserved between human and mouse.

Availability and implementation: xHeinz, an implementation of our algorithm, as well as all input data and results, are available at http://software.cwi.nl/xheinz and as a Galaxy service at http://services.cbib.u-bordeaux2.fr/galaxy in *CBiB Tools*.

Contact: gunnar.klau@cwi.nl
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Many computational methods have been proposed for the analysis of molecular profiles under different conditions. Studies employing these methods aim to better understand the molecular changes in the underlying cellular processes or to discover biomarkers as to classify between different conditions. Traditionally, analysis methods have been gene-centric, that is, they consider genes in isolation to establish differential patterns by simple statistical methods based on univariate statistical tests. For example, one of the first studies used gene expression measurements to differentiate between two leukemia classes (Golub *et al.*, 1999). With the availability of increasingly reliable biological network data for human and model organisms, gene-centric approaches have been increasingly complemented by integrative network analysis methods (Dittrich *et al.*, 2008; Ideker *et al.*, 2002; Mitra *et al.*, 2013). These methods yield *active modules*, that is, sets of genes that are connected in the network and show overall differential behavior. By taking the network topology into account, integrative analysis methods allow for a more robust interpretation of the measurements and result in more meaningful mechanistic insights.

Frequently, for ethical or practical reasons, molecular profiles are measured and validated on animal or cellular models and the results are therefore not immediately transferable to human (Okyere *et al.*, 2014). In fact, the low phase-II survival rate of 25% of potential drug compounds is largely attributed to the lack of transferability between model systems and human (Csermely *et al.*, 2013). This is also an issue in the recently discovered interleukin-17 producing helper T cells (Th17). These cells form a separate subset of helper T cells with a differentiation pathway distinct from those of the established Th1 and Th2 cells (Park *et al.*, 2005). Th17 cells are known to contribute to pathogenesis of inflammatory and autoimmune diseases such as asthma, rheumatoid arthritis, psoriasis and multiple sclerosis and play also a role in cancer immunology (Wilke *et al.*, 2011). Understanding the pathways and regulatory mechanisms that mediate the decision making processes resulting in the formation of Th17 is a critical step in the development of novel therapeutics. Unfortunately, the vast majority of data collected so far originates from studies performed on mice (Tuomela *et al.*, 2012) and, most importantly, a comprehensive comparison of the Th17 differentiation process in model organisms and in human is missing. Several studies indicate that the differentiation and phenotype of human and mouse Th17 cells are similar (Annunziato and Romagnani, 2009). Both subsets serve similar pro-inflammatory functions and produce the same hallmark cytokines and similar receptors. Furthermore, most of the already identified regulator genes show high sequence conservation. Other studies, however, show stimulus requirements for effective differentiation of human cells that differ from those required for mice (Annunziato *et al.*, 2009; McGeachy and Cua, 2008; O'Garra *et al.*, 2008). A characterization of the similarities and differences will not only increase our understanding of this fundamental process, but is also essential for sound translational research.

To do so, we suggest finding *conserved active modules* whose comprising genes show overall differential behavior, induce a connected subnetwork and are largely conserved across the species. Well-conserved modules make it possible to perform the experimental work and data analysis on the model organism. At the same time, the results are likely to be transferable to human. In addition, conserved modules carry a stronger signal than individual species modules because they integrate the signal of the individual data sources. Finding conserved active modules, however, is a difficult task. Separately computing species-specific active modules generally results in modules that are not conserved, which partially explains why experimental results are so often not transferable. Conversely, the largest conserved modules, as established, for example, with methods for network alignment, are not necessarily active. A computational model for finding conserved active modules requires thus a notion of both, activity and conservation (see Fig. 1).

Several authors already identified the benefits of combining and comparing cross-species experiments. At the single gene level, van Noort *et al.* (2003) have demonstrated that conserved co-expression is a strong co-evolutionary signal. More recent studies suggested to identify conserved biological processes. Lu *et al.* (2010) analyzed transcriptomics profiles of human and mouse macrophages and dendritic cells to derive common response genes involved in innate immunity. Kristiansson *et al.* (2013) proposed a method for the analysis of gene expression data that takes the homology structure between the different species into account. Berthier *et al.* (2012) found that murine and human responses to lupus nephritis involves similar gene networks. They first derived species-specific networks of significantly differentially expressed genes and then determined common subnetworks using a graph matching algorithm. Waltman *et al.* (2010) presented a multi-species integrative method to heuristically identify conserved biclusters. In their setting, a conserved bicluster is a subset of orthologous genes and a subset of conditions that achieve a high score with respect to co-expression, motif co-occurrence and network density. Dede and Oğul (2014) introduced a method that finds triclusters consisting of genes that are coexpressed across a subset of samples and a subset of species.

Deshpande *et al.* (2010) suggested the neXus algorithm for finding conserved active subnetworks. The authors use average fold

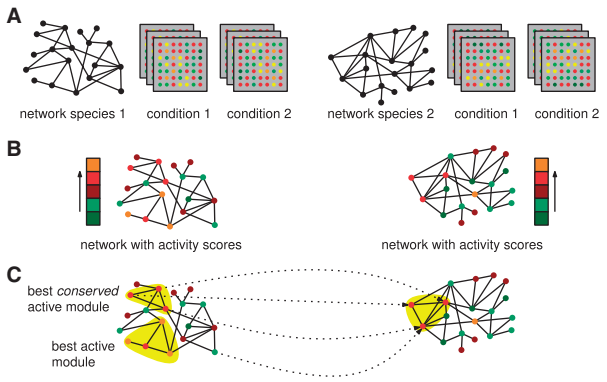

Fig. 1. Conserved active modules. Given two species-specific protein networks and, for each species, two sets of expression profiles of many different samples measured under two different conditions (**A**), we can annotate the nodes in the networks with activity scores (**B**), and identify modules that are at the same time highly differentially expressed and well-conserved (**C**). Cross-species conservation is indicated by dotted lines. Note that the best active module is not necessarily the best conserved active module

change of genes in a module as a measure for activity. To deal with conservation, they collapse paralogous genes within a cluster of orthologous genes (COG) (Tatusov *et al.*, 1997) into single nodes in the respective networks. They find modules using a seed-and-extend greedy heuristic that starts from a pair of orthologous seed nodes and then tries to simultaneously grow the two subnetworks by including pairs of neighboring orthologous genes. This strategy enforces a very stringent conservation policy: only modules whose genes are fully conserved are found. In addition, the locality of the greedy search strategy impairs the ability to find larger conserved modules and extending the search space around the seed genes drastically increases the runtime. In recent work, Zinman *et al.* (2015) introduce ModuleBlast, a method that, similarly to neXus, represents groups of orthologous proteins as single nodes in a combined network and tries to find connected subnetworks that are differentially expressed. The novelty of the method is the classification of the found modules according to the sign of the log fold change expression values. By doing so, the authors are able to assess whether conserved active modules show consistent or inconsistent expression patterns. Like neXus, ModuleBlast requires strict conservation of module genes.

Here, we propose a mathematical model for identifying conserved active modules for two species. It builds upon a model for single-species modules described in (Dittrich *et al.*, 2008) and inherits its notions for modularity and activity: A set of genes forms a module if it induces a connected subnetwork. The activity of a module is the sum of the activities of its genes, which are determined using a beta-uniform mixture model on the distribution of *P*-values that characterize the differential behavior. Instead of enforcing a stringent conservation policy, our model allows to specify the fraction of nodes in the solution that must be conserved. We cast our model as an integer linear programming formulation and present xHeinz, a branch-and-cut algorithm that, given enough time, solves this model to provable optimality, or, if stopped before, reports a solution with a quality guarantee. xHeinz is the first method that flexibly deals with conservation. We apply xHeinz to understand the mechanisms underlying Th17 T cell differentiation in both mouse and human. As a main biological result, we find that the key regulation factors of Th17 differentiation are conserved between human and mouse and demonstrate that all aspects of our model are needed to obtain this insight. We further demonstrate the robustness of our approach by comparing samples of the differentiation process obtained at different time points, in which we search for optimal, conserved active modules under a wide range of conservation ratios. Using a permutation test, we show that our results are statistically significant. Finally, we discuss the main differences between our results and the results obtained by the neXus tool on the same data set.

2 Approach

2.1 Mathematical model

We consider the conserved active modules problem in the context of two species networks, which we denote by $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$. Nodes in these networks are labeled by their activity—defined by $w \in \mathbb{R}^{V_1 \cup V_2}$ and conserved node pairs are given by the symmetric relation $R \subseteq V_1 \times V_2$. The aim is to identify two maximal-scoring connected subnetworks, one in each network, such that a given fraction α of module nodes are conserved. The formal problem statement is as follows:

PROBLEM 1 (Conserved active modules). *Given* $G_1 = (V_1, E_1)$, $G_2 = (V_2, E_2)$, $w \in \mathbb{R}^{V_1 \cup V_2}$ *and* $R \subseteq V_1 \times V_2$, *the task is to find a subset of nodes* $V^* = V_1^* \cup V_2^*$ *with* $V_1^* \subseteq V_1$ *and* $V_2^* \subseteq V_2$ *such that the following properties hold.*

- ***Activity:*** *Node activity scores are given by* $w \in \mathbb{R}^{V_1 \cup V_2}$, *where positive scores correspond to significant differential expression. For details see Section 3.2. We require that the sum* $\sum_{v \in V^*} w_v$ *is maximal.*
- ***Conservation:*** *Conserved node pairs are given by the relation* $R \subseteq V_1 \times V_2$. *We require that at least a certain fraction* α *of the nodes in the solution must be conserved, that is,* $|U^*| \geq \alpha \cdot |V^*|$ *where* $U^* := \{u \in V_1^* | \exists v \in V_2^* : uv \in R\} \cup \{v \in V_2^* | \exists u \in V_1^* : uv \in R\}$.
- ***Modularity:*** *We require that the induced subgraphs* $G_1[V_1^*]$ *and* $G_2[V_2^*]$ *are connected.*

The model allows a trade-off between conservation and activity. If no conservation is enforced ($\alpha = 0$), the solution will correspond to two independent maximum-weight connected subgraphs. Conversely, if complete conservation is required ($\alpha = 1$), the solution can only consist of conserved nodes, which results in lower overall activity. The user controls this trade-off by varying the value of the parameter α from 0 to 1. The activity score monotonically decreases with increasing α (see Fig. 2).

Since the maximum-weight connected subgraph problem, which occurs as a subproblem for $\alpha = 0$, is NP-hard (Johnson, 1985), the problem of finding conserved active modules is NP-hard as well.

2.2 Integer linear programming approach

We formulate the conserved active modules problem as an integer programming (IP) problem in the following way.

$$\max \sum_{v \in V_1 \cup V_2} w_v x_v \tag{1}$$

$$\text{s.t. } m_u = \max_{uv \in R} \{x_u x_v\} \qquad u \in V_1 \tag{2}$$

$$m_v = \max_{uv \in R} \{x_u x_v\} \qquad v \in V_2 \tag{3}$$

$$\sum_{v \in V_1 \cup V_2} m_v \geq \alpha \sum_{v \in V_1 \cup V_2} x_v \tag{4}$$

$$G_1[\mathbf{x}] \text{ and } G_2[\mathbf{x}] \text{ are connected} \tag{5}$$

$$x_v, m_v \in \{0, 1\} \qquad v \in V_1 \cup V_2 \tag{6}$$

Variables $\mathbf{x} \in \{0,1\}^{V_1 \cup V_2}$ encode the presence of nodes in the solution, i.e. for all $v \in V_1 \cup V_2$ we want $x_v = 1$ if $v \in V^*$ and $x_v = 0$

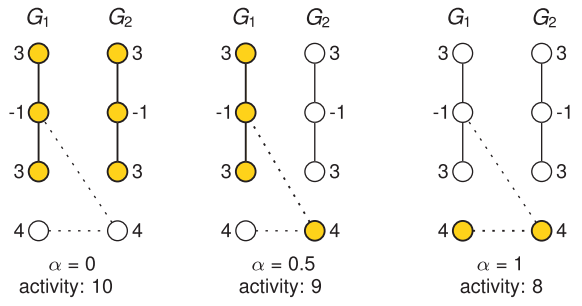

Fig. 2. Trade-off between activity and conservation. Three optimal solutions (indicated in yellow) for varying conservation ratios α in a toy example instance. Node activities are given next to the nodes, conserved node pairs are linked by dotted lines. The activity of a conserved module is the sum of the activities of its comprising nodes. The parameter α denotes the minimum fraction of nodes in a solution that must be conserved, i.e. connected by a dotted line

otherwise. The objective function (1) uses these variables to express the activity of the solution, which we aim to maximize. Variables $\mathbf{m} \in \{0,1\}^{V_1 \cup V_2}$ encode the presence of conserved nodes in the solution. Constraints (2) encode that a node $u \in V_1$ that is present in the solution ($x_u = 1$) is conserved if there exists a related node $v \in V_2$ ($uv \in R$) that is also present in the solution ($x_v = 1$). Similarly, constraints (3) define conserved nodes in V_2 that are present in the solution. The fraction of conserved nodes in the solution is at least α as captured by Constraint (4). In addition, we satisfy the modularity property by requiring in Constraint (5) that $G_1[\mathbf{x}]$ and $G_2[\mathbf{x}]$ are connected. In Supplementary Text A.1 we give further details on how to model Constraints (2), (3) and (5) as linear inequalities and on the implementation that solves this formulation.

3 Material and methods

3.1 Experimental procedure

We summarize here the experimental procedure followed by Tuomela *et al.* (2012) and Yosef *et al.* (2013) to generate transcriptomic profiles. In (Tuomela *et al.*, 2012), CD4 + T-cells were isolated from umbilical cord blood of several healthy neonates, arranged in three different pools, then activated with anti-CD3 and anti-CD28. Cells from each pool were then divided in two batches, one to be polarized toward Th17 direction, and one serving as control (Th0). Th17 differentiating cytokines consisted of IL6 (20 ng/mL), IL1B (10 ng/mL) and TGFB (10 ng/mL), along with neutralizing anti-IFNG (1 μg/mL) and anti-IL4 (1 μg/mL). Three biological replicates of human cells, for both conditions (coming from each pool), were collected between 0.5 and 72 h (0.5, 1, 2, 4, 6, 12, 24, 48, 72 h time points) and hybridized on Illumina Sentrix HumanHT-12 Expression BeadChip Version 3. The microarray data were analyzed using the beadarray Bioconductor package (Dunning *et al.*, 2007). In (Yosef *et al.*, 2013), CD4 + T-cells were purified from spleen and lymph nodes from wild type C57BL/6 mice, then activated with anti-CD3 and anti-CD28. For Th17 differentiation, cells were cultured with TGFB (2 ng/mL), IL6 (20 ng/mL), IL23 (20 ng/mL) and IL1B (20 ng/mL) during 0.5–72 h (at time points 0.5, 1, 2, 4, 6, 8, 10, 12, 16, 20, 24, 30, 42, 48, 50, 52, 60, 72 h), and finally hybridized on an Affymetrix HT_MG-430 A.

3.2 Microarray processing, statistical analysis and node scoring

Preprocessed and quantile normalized data sets were downloaded from GEO under the accession numbers GSE43955 and GSE35103. As downloaded from GEO, both the human and the mouse time-series were already filtered by retaining only the probes with detection P-values < 0.05 in at least one time point and one condition. Following the original studies, we further only retained probes having a standard deviation > 0.15 over all the conditions and time points; as well as being annotated by a single ENSEMBL gene. Finally, a single probe was selected for each gene by taking, for each ENSEMBL gene, the probe having the largest variance accross all samples. In total, 12 307 and 18 497 probes passed the filters for the mouse and human data set, respectively.

Differential expression between Th17 and Th0 conditions were estimated using the limma package (Smyth, 2005). Human samples were indicated as paired according to the experimental design so as to account for the pooled human samples. For mouse samples, calling was performed on all Th0 versus Th17 samples, regardless of the mouse donor. To determine which genes were differentially expressed at a given time point, we used a linear model to estimate the interaction between the treatment and the time effect. The linear models used for the human and mouse studies include one interaction term for each time point and exclude the intercept (In R, the formula reads: $\sim 0 + \text{treat} : \text{time}$). Differential expression at any time point K of interest were determined by the contrasts $\text{Th17.time}_K - \text{Th0.time}_K$. We report in this study results for the following time points: 2, 4, 24, 48, 72 h.

Following (Dittrich *et al.*, 2008), we computed positive and negative scores for each gene at each time point by fitting a beta-uniform mixture model using the implementation in the BioNet package (Beisser *et al.*, 2010). For a detailed description of this procedure, see Supplementary Text A.2. Throughout this study, FDR = 0.1 was used for all samples and species.

Due to the experimental noise and paired design, the human samples have much higher intra-group variance, resulting in significant calls having p-values orders of magnitude higher than the mouse calls. This results in a range of scores that is much narrower for human than for mouse, possibly imbalancing results towards mouse modules. To correct for this effect, scores of mouse genes were rank normalized to the scores of the human genes as follows: the scores were sorted, and for each gene the score of the ith mouse gene was set to the score of the ith human gene. Comparison of the distribution of scores before and after normalization showed that compared to usual Benjamini-Hochberg FDR and log fold change cut-offs ($|\log \text{FC}| \geq 1$), the loss in statistical power was inconsequential and that this procedure ensured that mouse and human genes had comparable score distributions.

3.3 Network and orthology databases

The human and mouse background networks were downloaded from STRING v9.1, protein.actions.detailed.v9.1.txt (Franceschini *et al.*, 2013), which is a database that contains experimentally verified direct protein interactions. Note that this network also contains interactions predicted based on orthology, so-called *interologs*. Ideally, we would prefer to use only experimentally predicted interactions, but currently, for mouse, such available data is too incomplete to result in a meaningful background network. Outlier nodes with a degree above 40 times the interquartile range plus the 75th percentile of the distribution of all node degrees were removed (ELAVL1, UBC, Ubb, Ubc). The resulting mouse network has 16 821 nodes and 483 532 edges and the human network has 16 255 nodes and 315 442 edges.

Orthology information was downloaded from Ensembl release 59 (Flicek *et al.*, 2013) and all human and mouse orthologs were kept, regardless of the identity scores. The orthology mapping corresponds to a bipartite graph involving 67 304 human proteins and 43 953 mouse proteins linked by 104 007 edges, grouped in 16 552 bicliques with an average size of 6.72 proteins (SD: 5.34).

3.4 Implementation, input and output

xHeinz is implemented in modern C++, using the boost libraries and the LEMON graph library (Dezső *et al.*, 2011). CPLEX 12.6 is used to solve the ILP. The source code is publicly available in a git repository linked to from http://software.cwi.nl/xheinz.

xHeinz takes as input (i) two species-specific networks, (ii) an orthology mapping between the nodes of the two networks, (iii) scores associated to each of the nodes, e.g. derived from the P-value of the moderated t-test, (iv) the threshold value α and (v) an optional time limit.

We performed a preprocessing step where we retained the subgraphs of the input networks induced by the genes that meet the microarray filtering criteria. This reduced the number of nodes to 8453 human nodes, 6882 mouse nodes and 14 779 nodes in the orthology mapping. Among these, up to 250 nodes (depending on the time point) have positive scores. The rank normalization as described in Section 3.2 ensured that the number of positive human nodes is in the order of the number of positive mouse nodes.

xHeinz returns two node sets corresponding to a solution found within the time limit together with an upper bound on the optimal solution value. In case the solution value equals this upper bound, the computed solution is provably optimal.

4 Results and discussion

4.1 xHeinz identifies conserved modules at different levels of conservation

We applied xHeinz on samples from the Th17 human and mouse data sets for time points 2, 4, 24, 48 and 72 h. We solved these instances for different values of $\alpha \in [0, 1]$ with a step size of 0.1. All computations were done in single-thread mode on a desktop computer (Intel XEON e5 3 Ghz) with 16 Gb of RAM and a time limit of 12 000 CPU seconds. After this timeout, the best feasible solution is returned by the solver.

Figure 3 shows for the five time points and eleven values of the α parameter, the human and mouse scores of the found modules as well as the distribution of the module contents. For 26 of the 55 instances we solved the conserved active modules problem to provable optimality within the time and memory limit. The optimality gap of a solution is defined as $(UB - LB)/|LB|$, where LB and UB are the value of the best solution and the lowest upper bound as identified by the branch-and-cut algorithm, respectively. Of the 29 instances that are not solved to optimality, 22 have a gap smaller than 5%.

Any feasible solution for a conservation ratio of α is also a solution for any $\alpha' \leq \alpha$. We indeed see in Figure 3 that this property holds, the solution values decrease monotonically with increasing α. Also the solutions for $\alpha = 0$ (no conservation constraints) are identical to the solutions obtained by running the single species method Heinz (Dittrich *et al.*, 2008) separately on the two networks.

There is a sharp decrease in module size for $\alpha = 1$. Indeed, this is the most restrictive setting since it enforces that all the nodes in a module must be conserved. We also observe that as α increases, both positive and negative *conserved* nodes are added, indicating that we manage to retrieve informative nodes in a gradual manner. See also Supplementary Text A.8 for a detailed analysis of module overlap for all combinations of α values.

When we compare solutions across time points, we see that the conserved active modules capture two phases of the differentiation process. We observe high activity at 2 h as well as at the late time points. Several authors reported such biphasic behavior during early Th17 differentiation, both in mouse (Ciofani *et al.*, 2012; Yosef *et al.*, 2013) and human (Tuomela *et al.*, 2012). The low activity score observed at the 4 h time point is in line with previous mouse studies, which suggest that after the initial induction sustained by Stat3 and Stat1 in the first four hours, a phase of Rorc induction takes place and lasts until the 20 h time point, after which the effective protein level of Rorc starts to increase and to trigger the cytokine production phase (Yosef *et al.*, 2013). Our model and the solutions obtained suggest that these dynamics are conserved between the two organisms.

4.2 Early regulation of Th17 differentiation is conserved between human and mouse

In the following, we study the two phases of the Th17 differentiation process in more detail. We focus on the 2 and 48 h time points. We selected for this evaluation $\alpha = 0.8$ for both time points, as this value provides a balance between conservation and activity and

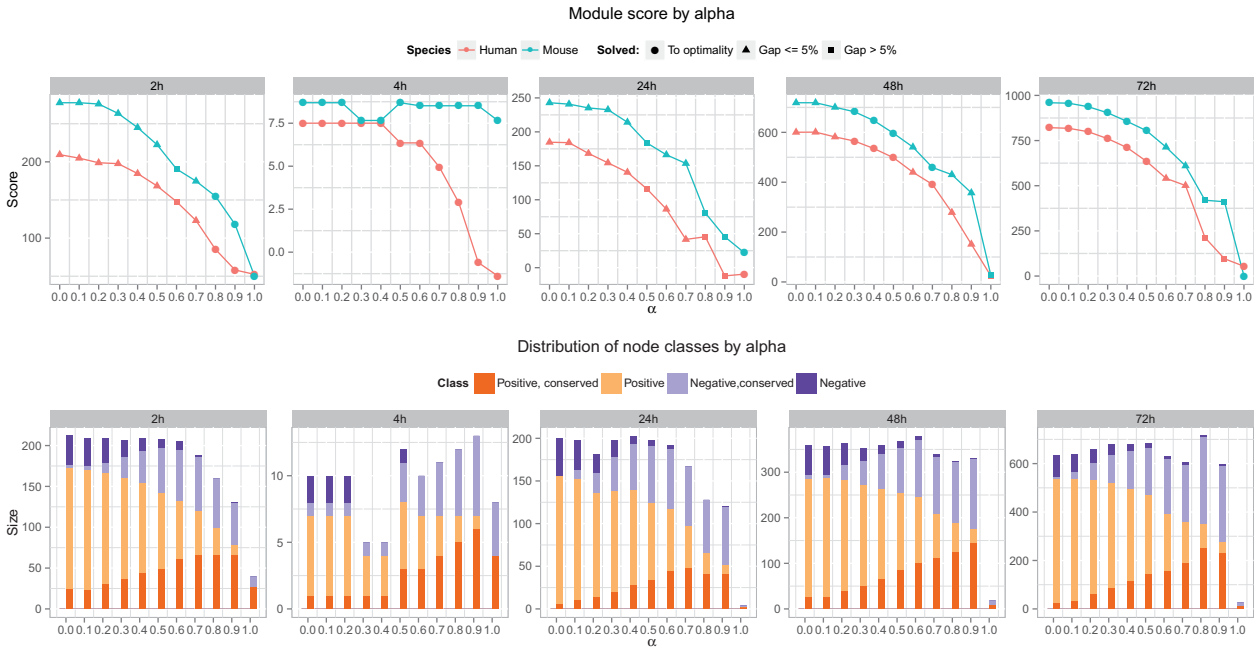

Fig. 3. Statistics of xHeinz solutions. The conserved active module problem was solved for five time points (columns) over a sequence of 11 consecutive values of the α conservation parameter (x-axis). We report in the top row the score of the best solution (y-axis) and whether optimality was proven by our algorithm (circles). The second row illustrates how module contents vary as α increases. The height of each bar indicates the size of the respective module, colors indicate the fraction of positive and conserved nodes

produces modules of interpretable size. All results at all time points are available on the accompanying website. Figure 4 reports the resulting human and mouse modules for the two time points.

We assess statistical significance of the resulting modules by performing 100 runs on randomized networks for each value of α, and additional 400 runs for the selected $\alpha = 0.8$. We do this using two randomization methods: (i) permuting the node weights while keeping the graph fixed, and (ii) permuting the network topology while keeping the node weights and the node degrees fixed as described in Mihail and Zegura (2003). With the exception of a few extreme cases at the 48 h time point, all modules were found to be highly significant. For details see Supplementary Text A.8.

At the 2 h time point, xHeinz identifies a conserved module consisting of 58 human and 50 mouse proteins. Interestingly, both the human and mouse modules are centered around STAT3/Stat3. STAT3 is a signal transducer having transcription factor activity and was shown to play a key role in the differentiation process of Th17 (Harris *et al.*, 2007). Once activated by Th17 polarizing cytokines (such as IL6 in our case), it eventually binds to the promoter regions of IL17A/Il17a and IL17F/Il17f cytokines and activates transcription. These cytokines are the hallmark cytokines produced by activated Th17 cells. It is worth noting that IL17/Il17 cytokines and associated receptors are not in the 2 h modules, as these proteins have been shown to be expressed only at later time points (Tuomela *et al.*, 2012). Moreover, STAT1/Stat1, another member of the STAT family, is part of the solution and belongs to the central core of the human and mouse modules, which is consistent with its major role during the early phases of Th17 differentiation (Yosef *et al.*, 2013).

We also observe that the STAT3/BATF/IL6ST/SOCS3 region of the 2 h module is well-conserved. Batf has been shown to directly control Th17 differentiation in mouse (Schraml *et al.*, 2009) and BATF proteins are detected as early as after 12 h of polarization in human (Tuomela *et al.*, 2012). Similarly, SOCS3 is a known IL6 and IL21-induced negative regulator of Th17 polarization, that is eventually down-regulated by TGFB and IL6ST at a later phase in order to prolong STAT3 activation (Qin *et al.*, 2009; Zhu *et al.*, 2008). Overall, these modules show highly conserved and significant enrichment for response to cytokine stimulus (Benjamini-Hochberg (BH) FDR 5.6 e−4), JAK-STAT (BH FDR 4.8 e−4) cascade and transcription regulator activity (BH FDR 2.3 e−4), computed using the DAVID functional annotation chart (Huang *et al.*, 2009). This indicates that the identified module matches expected biological mechanisms observed at early phases (Ciofani *et al.*, 2012). Furthermore, comparison of the dynamics of expression shows that genes differentially expressed in both species change expression in the same direction (cf. Supplementary Text A.3).

We also applied xHeinz to find a conserved module at a later time point (48 h). Kinetics analysis of Th17 differentiation showed that the effective secretion of Th17 hallmark cytokines only happens after several days of polarization (Tuomela *et al.*, 2012; Yosef *et al.*, 2013) and we do observe in these modules a significant enrichment for interleukin related proteins present in both species, which was absent for the 2 h modules, such as up-regulation of IL9/Il9. Secretion of IL9 by Th17 cells have been demonstrated both in mouse and human cells (Beriou *et al.*, 2010), Il9 is know to be induced by Bcl3 (Richard *et al.*, 1999), and Bcl3 inhibition has been recently shown to affect the function of Th17 cells in mouse

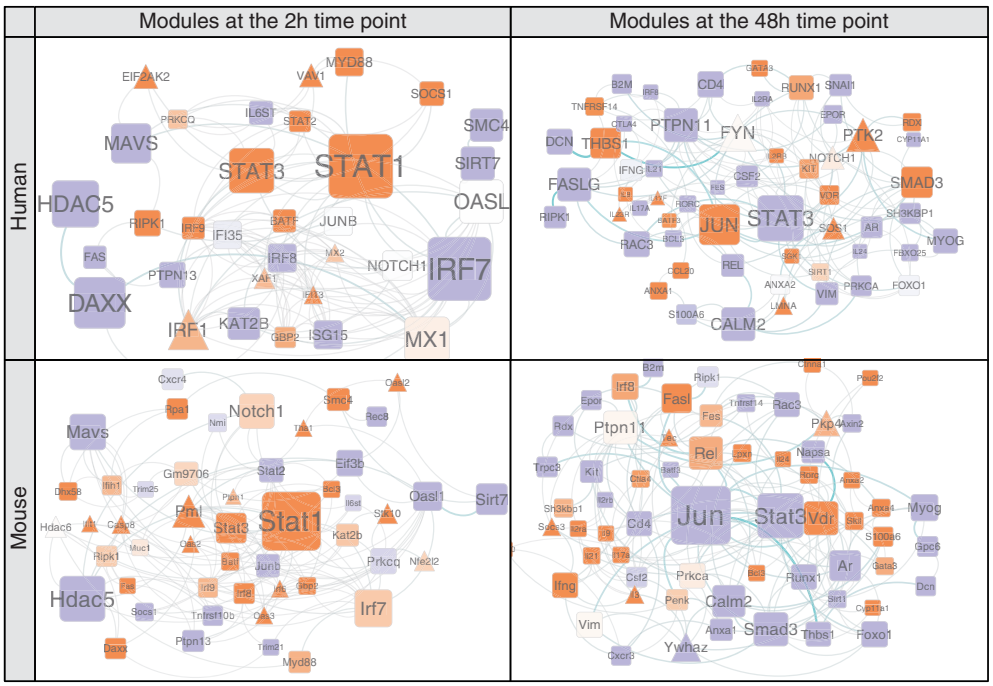

Fig. 4. Conserved active Th17 differentiation modules in human and mouse at 2 and 48 h. We obtained node activity scores capturing the significance of differential gene expression between the Th17 and Th0 conditions in human and mouse using the BUM model with FDR = 0.1. xHeinz uses these scores to search for conserved active modules in the STRING protein action network. The first row shows the human counterparts of the best scoring conserved modules for the 2 h (left) and 48 h (right) samples. The second row depicts the mouse counterparts. Rounded squares depict genes for which a homolog—as defined by Ensembl—is present in the counterpart, whereas triangles denote non-conserved genes. Node color gradually indicates activity scores. Orange: larger than 2; white: between –2 and 2; violet: smaller than –2. Node labels and sizes are proportional to betweenness centrality and edge width to edge-betweenness—both centralities are with respect to the subnetwork module. Only nodes having a degree larger than 2 (resp. 3) are displayed for the 2 h (resp. 48 h) module. The full networks are available on the accompanying website and in Supplementary Text A.3

(Ruan *et al.*, 2010). We also observe the conserved down-regulation of GATA3/Gata3, which is known to be the master regulator of Th2 cells (Zheng and Flavell, 1997), and is likely to constrain the Th17 regulation program (van Hamburg *et al.*, 2008). Similarly to the modules found at 2 h, the 48 h modules are centered around STAT3, although at the 48 h time point this gene is not differentially expressed anymore neither in human or mouse (resp. logFC of 0.17, score of −4.59 for human, and logFC 0.52, score of −3.21 for mouse). This observation is in line with the major role of STAT3 along the differentiation process at all time points (Yosef *et al.*, 2013). To the contrary, STAT1 has been indicated as an exclusively early regulator (Yosef *et al.*, 2013) in mouse and is indeed not present anymore in the 48 h modules. We also observe the presence of the RORA/RORC/Rora/Rorc members of the RORs family of intracellular transcription factors, which are considered to be the master regulators of the Th17 lineage (Yang *et al.*, 2008), and have been implicated in both species (Crome *et al.*, 2009). Interestingly, these regulators are linked to the up-regulation of the vitamin-D receptor (VDR/Vdr), whose role in Th17 differentiation and several human auto-immune related disease have been recently studied (Chang *et al.*, 2010).

In summary, our findings show the relevance of the identified conserved active modules with regard to the biological process of interest. By requiring the active modules to contain a certain fraction of conserved nodes, xHeinz identifies the main core proteins involved in the differentiation of Th17. Our analysis confirms that these proteins are very likely to have similar roles in both species.

4.3 Comparison to neXus

We compare the 48 h xHeinz modules (*cf.* Fig. 4) with subnetworks computed by neXus version 3 (Deshpande *et al.*, 2010). neXus uses a heuristic technique to grow subnetworks from seed nodes simultaneously in two species. This is done in an iterative fashion. Neighborhoods of the two current modules are determined using a depth-first search. This search is restricted to only consider nodes that have a path to the seed node with a confidence larger than the user-specified parameter `dfscutoff`. The confidence of a path is defined as the product of the confidences of the edges comprising that path. The modules are extended to include the most active pair of orthologous nodes in the neighborhoods—where activity is defined as normalized log fold change and thus differs from the definition of activity used in xHeinz. This whole procedure is repeated until either the cluster coefficient drops below the user-specified parameter `cc`, or the average activity scores of one of the two modules drops below parameter `scorecutoff`. We ran neXus with the default parameters `cc` = 0:1, 0:2, `scorecutoff` = 0:15 and `dfscutoff` = 0:3, 0:8 for mouse and human respectively for all time points. Table 1 gives the resulting module sizes for human and mouse.

neXus finds 1 module for time point 48 h which is shown in Figure 5 for human (A) and mouse (B). In total 5 genes are contained in the module, which are identical for human and mouse, but the number of edges differs. Only one of the genes is significantly differentially expressed, CCL20, which has an absolute log fold change bigger than 1 and a BH FDR smaller than 0.1. Since neXus does not use p-values as an input, but log fold-changes which are normalized to activity values, the genes CCL20 and CXCR3 are considered as active nodes with a value above 0.15. These genes show changes in expression, but only two of these changes are statistically significant. The low number of active nodes points to a drawback in the neXus algorithm: due to the locality of the greedy search strategy it may happen that the average activity of the subnetwork in construction keeps on degrading without reaching the next active node. The effects of this issue can be seen, for example, in Figure 5, where CCL20 is the seed node and the majority of other neighboring nodes are not differentially expressed.

Another consequence of the neXus search strategy is that the module sizes are small (*cf.* Table 1) and thus only give a limited view of the molecular mechanisms at play. Theoretically, the parameter `dfscutoff` can be decreased to increase the module size. Doing so, however, produces only slightly larger modules, but drastically increases the running time (Supplementary Table S1). Changes in the clustering coefficient parameter cc only reduce the module size with increasing cc (Supplementary Table S2).

Conservation in neXus is enforced stringently by only allowing pairs of orthologous genes or genes that are only present in one of the networks to be included in the subnetworks (see Fig. 5). This is too restrictive if the underlying mechanisms in the two species differ. For instance, for time point 48 h and all but $\alpha = 1$ values, xHeinz finds the non-conserved gene IL23R (BH FDR 3.52 e−8, score 14.50, logFC 1.38) in human, which is involved in Th17 autocrine signaling (Wei *et al.*, 2007) but which is not differentially expressed in mouse. xHeinz also finds JUNB, which at the 2 h time point is up-regulated in human data (BH FDR 1 e−2, score 0.02, logFC 1.3) and not detected as differentially expressed in the mouse data (BH

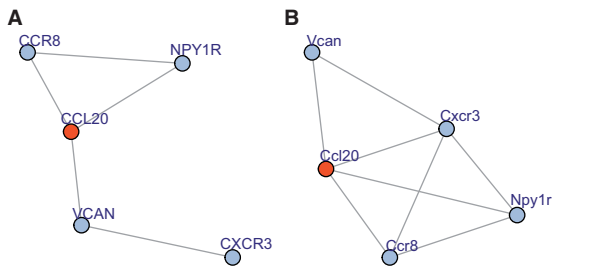

Fig. 5. neXus module for the time point 48 hours for human (**A**) and mouse (**B**). Orange coloring indicates genes with significant differential expression (BH FDR ≤ 0.1, $|\log FC| \geq 1$). Here only one gene is significantly differentially expressed (CCL20)

Table 1. Modules calculated with neXus for all time points

Solution	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	avg.	#sols
0.5 h	7 (6)	4 (4)	7 (6)	3 (3)												5.25 (4.75)	4
1 h	15 (10)	10 (9)	12 (12)	13 (13)	7 (7)	5 (5)	15 (13)	10 (11)	9 (10)	18 (16)	25 (24)	14 (14)	6 (7)	5 (5)	6 (6)	9.95 (9.58)	19
2 h	15 (17)	6 (5)	12 (10)	12 (11)	10 (10)	13 (13)	17 (15)	12 (12)	8 (9)	5 (5)	11 (12)	19 (18)	3 (3)	9 (8)	23 (21)	10 (9.83)	30
4 h	6 (9)	4 (4)	6 (5)	4 (4)	4 (4)	3 (3)	7 (8)	9 (10)	4 (4)	3 (3)						5 (5.40)	10
48 h	5 (5)															5 (5)	1

Note: Shown are the sizes in number of nodes of the first 15 representative solutions and the average sizes for the human subnetwork and for the mouse subnetwork in brackets. The last column lists the number of solutions for each time point. No solutions were obtained for time points 24 and 72 h.

FDR 0.48, score −4.01, logFC 0.65). JUNB is a known partner of BATF with which it heterodimerizes preferentially during Th17 differentiation (Schraml *et al.*, 2009), indicating its relevance. Both important genes would have been missed by a more restrictive conservation setting. Indeed, both neXus and xHeinz at $\alpha = 1$ fail to find these genes showing that a more flexible view on conservation is required to adequately deal with transferability.

5 Conclusion

We introduce a mathematical model for the problem of finding active subnetwork modules that are conserved between two species and thus contribute to formalizing the notion of conserved active modules. A key feature of our model is a flexible notion of conservation, which is controlled by a parameter $\alpha \in [0, 1]$: We require that at least a fraction α of the nodes are conserved between the species-specific modules of a solution. Note that in case of more distantly-related species a smaller α value may be more appropriate. We have translated our model into an integer linear programming formulation and have devised and implemented an exact branch-and-cut algorithm that computes provably optimal or near-optimal conserved active modules in our model.

Our computational experiments for understanding the mechanisms underlying Th17 T cell differentiation in both mouse and human demonstrate that the flexibility in the definition of conservation is crucial for the computation of meaningful conserved active modules. We have found two conserved Th17 modules at time points 2 h ($\alpha = 0.8$) and 48 h ($\alpha = 0.8$) that thoroughly encompass the biphasic Th17 differentiation process. This result can not be revealed by requiring full conservation ($\alpha = 1$) or by independent modules without requiring conservation ($\alpha = 0$). Likewise, neXus, an alternative approach based on a stringent conservation model, is not able to capture the key regulatory program of the differentiation process.

A key characteristics of our model is its flexibility. This allows its extension to multiple species and time points, which we will address in future work. In this case, however, realistic instances will be harder to compute to optimality and will require the development of powerful algorithm engineering techniques.

Acknowledgement

We thank the three anonymous referees for their constructive comments.

Funding

This work was supported in part by the SIRIC BRIO grant (Site de Recherche Intégrée sur le Cancer-Bordeaux Recherche Intégrée Oncologie) to HS. Computer resources were provided by the Bordeaux Bioinformatics Center (CBiB), Université de Bordeaux.

Conflict of Interest: none declared.

References

Annunziato,F. and Romagnani,S. (2009) Do studies in humans better depict Th17 cells? *Blood*, **114**, 2213–2219.

Annunziato,F. *et al.* (2009) Human Th17 cells: are they different from murine Th17 cells? *Eur. J. Immunol.*, **39**, 637–640.

Beisser,D. *et al.* (2010) BioNet: an R-package for the functional analysis of biological networks. *Bioinformatics*, **26**, 1129–30.

Beriou *et al.* (2010) TGF-beta induces IL-9 production from human Th17 cells. *J. Immunol.*, **185**, 46–54.

Berthier,C.C. *et al.* (2012) Cross-species transcriptional network analysis defines shared inflammatory responses in murine and human lupus nephritis. *J. Immunol.*, **189**, 988–1001.

Chang,S.H. *et al.* (2010) Vitamin D suppresses Th17 cytokine production by inducing C/EBP homologous protein (CHOP) expression. *J. Biol. Chem.*, **285**, 38751–38755.

Ciofani,M. *et al.* (2012) A validated regulatory network for Th17 cell specification. *Cell*, **151**, 289–303.

Crome,S.Q. *et al.* (2009) The role of retinoic acid-related orphan receptor variant 2 and IL-17 in the development and function of human CD4+ T cells. *Eur. J. Immunol.*, **39**, 1480–1493.

Csermely,P. *et al.* (2013) Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol. Therap.*, **138**, 333–408.

Dede,D. and Oğul,H. (2014) TriClust: a tool for cross-species analysis of gene regulation. *Mol. Inf.*, **33**, 382–387.

Deshpande,R. *et al.* (2010) A scalable approach for discovering conserved active subnetworks across species. *PLoS Comput. Biol.*, **6**, e1001028.

Dezső,B. *et al.* (2011) LEMON–an open source C++ graph template library. *Electr. Notes Theor. Comput. Sci.*, **264**, 23–45.

Dittrich,M.T. *et al.* (2008) Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics (Oxford, England)*, **24**, i223–i231.

Dunning,M.J. *et al.* (2007) beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics*, **23**, 2183–2184.

Flicek,P. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–D55.

Franceschini,A. *et al.* (2013) STRING v9.1: protein–protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–D815.

Golub,T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Harris,T.J. *et al.* (2007) Cutting edge: an in vivo requirement for STAT3 signaling in TH17 development and TH17-dependent autoimmunity. *J. Immunol.*, **179**, 4313–4317.

Huang,D.W. *et al.* (2009) Systematic and integrative analysis of large gene lists using David bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Ideker,T. *et al.* (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics (Oxford, England)*, **18**, S233–S240.

Johnson,D.S. (1985) The NP-completeness column: an ongoing guide. *J. Algorithms*, **6**, 145–159.

Kristiansson,E. *et al.* (2013) A novel method for cross-species gene expression analysis. *BMC Bioinformatics*, **14**, 70.

Lu,Y. *et al.* (2010) Cross species expression analysis of innate immune response. *J. Comput. Biol.*, **17**, 253–68.

McGeachy,M.J. and Cua,D.J. (2008) Th17 cell differentiation: the long and winding road. *Immunity*, **28**, 445–453.

Mihail,C.G.M. and Zegura,E. (2003) The markov chain simulation method for generating connected power law random graphs. In: *Proceedings of the Fifth Workshop on Algorithm Engineering and Experiments*. Vol. 111, SIAM, pp. 16.

Mitra,K. *et al.* (2013) Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.*, **14**, 719–732.

O'Garra,A. *et al.* (2008) Differentiation of human T(H)-17 cells does require TGF-beta! *Nat. Immunol.*, **9**, 588–590.

Okyere,J. *et al.* (2014) Cross-species gene expression analysis of species specific differences in the preclinical assessment of pharmaceutical compounds. *PLoS ONE*, **9**, e96853.

Park,H. *et al.* (2005) A distinct lineage of CD4 T cells regulates tissue inflammation by producing interleukin 17. *Nat. Immunol.*, **6**, 1133–1141.

Qin,H. *et al.* (2009) TGF-beta promotes Th17 cell development through inhibition of SOCS3. *J. Immunol.*, **183**, 97–105.

Richard,M. *et al.* (1999) Interleukin-9 regulates NF-kappaB activity through BCL3 gene induction. *Blood*, **93**, 4318–4327.

Ruan,Q. *et al.* (2010) Roles of bcl-3 in the pathogenesis of murine type 1 diabetes. *Diabetes*, **59**, 2549–2557.

Schraml,B.U. *et al.* (2009) The AP-1 transcription factor batf controls T(H)17 differentiation. *Nature*, **460**, 405–409.

Smyth,G.K. (2005) Limma: linear models for microarray data. In: Gentleman,R. *et al*. (eds.) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Springer, New York, pp. 397–420.

Tatusov,R. *et al*. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Tuomela,S. *et al*. (2012) Identification of early gene expression changes during human Th17 cell differentiation. *Blood*, **119**, e151–e160.

van Hamburg,J.P. *et al*. (2008) Enforced expression of GATA3 allows differentiation of IL-17-producing cells, but constrains Th17-mediated pathology. *Eur. J. Immunol.*, **38**, 2573–2586.

van Noort,V. *et al*. (2003) Predicting gene function by conserved co-expression. *Trends Genet.*, **19**, 238–242.

Waltman,P. *et al*. (2010) Multi-species integrative biclustering. *Genome Biol.*, **11**, R96.

Wei,L. *et al*. (2007) IL-21 is produced by Th17 cells and drives IL-17 production in a STAT3-dependent manner. *J. Biol. Chem.*, **282**, 34605–34610.

Wilke,C.M. *et al*. (2011) Deciphering the role of Th17 cells in human disease. *Trends Immunol.*, **32**, 603–611.

Yang,X.O. *et al*. (2008) T helper 17 lineage differentiation is programmed by orphan nuclear receptors ROR alpha and ROR gamma. *Immunity*, **28**, 29–39.

Yosef,N. *et al*. (2013) Dynamic regulatory network controlling TH17 cell differentiation. *Nature*, **496**, 461–468.

Zheng,W. and Flavell,R.A. (1997) The transcription factor gata-3 is necessary and sufficient for th2 cytokine gene expression in cd4 t cells. *Cell*, **89**, 587–596.

Zhu,B.-M. *et al*. (2008) SOCS3 negatively regulates the gp130-STAT3 pathway in mouse skin wound healing. *J. Invest. Dermatol.*, **128**, 1821–1829.

Zinman,G.E. *et al*. (2015) ModuleBlast: identifying activated sub-networks within and across species. *Nucleic Acids Res.*, **43**, e20–e20.