Journal of Immunological Methods 387 (2013) 71–80

Contents lists available at SciVerse ScienceDirect

Journal of Immunological Methods

journal homepage: www.elsevier.com/locate/jim

Research paper

Decoding dendritic cell function through module and network analysis ☆

Gaurav Pandey [a,*], Ariella Cohain [a], Jennifer Miller [b], Miriam Merad [b]

[a] Institute for Genomics and Multiscale Biology and Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY, USA
[b] The Immunology Institute, Tisch Cancer Institute and Department of Oncological Sciences, Mount Sinai School of Medicine, New York, NY, USA

ARTICLE INFO

Article history:
Received 17 September 2012
Accepted 17 September 2012
Available online 23 October 2012

Keywords:
Dendritic cells
Gene expression data
Clustering
Network analysis
Transcription factors
Regulatory networks

ABSTRACT

Systems biology approaches that utilize large genomic data sets hold great potential for deciphering complex immunological process. In this paper, we propose such an approach to derive informative modules and networks from large gene expression data sets. Our approach starts with the clustering of such data sets to derive groups of tightly co-expressed genes, also known as co-expression modules. These modules are then converted into co-expression networks, and combined with transcriptional regulatory and protein interaction data to generate integrated networks that can help decipher the regulatory structure of these modules. We use this approach to derive the first set of modules and networks focused on dendritic cells (DCs). These cells are responsible for sampling the local environment to inform the adaptive immune system about peripheral stimuli, thus leading to the induction of an immune response. Using the ImmGen gene expression data set, we derive co-expression modules and integrated networks for the pDC, cDC and CD8+ DC subsets. In addition to recapitulating genes known to regulate the functions of these subsets, these networks reveal several novel genes and interactions that might have important roles in DC biology.

1. Introduction

Immune responses result from a complex interaction that relies on an elaborate and dynamic communications network that exists among the many different immune cell types that patrol the body. Although several of the cellular and molecular cues that control the induction of successful immune responses have been identified, there is an immense need for a systems-level understanding of how the different components of immune cells interact in the steady state and in response to different stimuli. To address this need, several groups have started utilizing the recent wave of biotechnologies to profile the immune system at the molecular level and analyze the related data to obtain novel insights (Gardy et al., 2009; Germain et al., 2011). In particular, the ImmGen consortium has profiled the genome-wide expression patterns in all the cell types in the immune system of *Mus musculus* (mouse), thus making available an unprecedented resource for such studies (Heng and Painter, 2008). This and other data sets have been utilized by some recent rigorous computational systems biology approaches that have built models of how the different components of the immune system function individually and in concert with the others (Amit et al., 2009; Germain et al., 2011; Novershtern et al., 2011; Benichou et al., 2012). However, the findings of these studies have largely been limited to the immune cells where rich data sets are available, such as T- and B-cells.

An important component of the immune system whose understanding has not benefitted much from these studies is dendritic cells (DCs) (Banchereau and Steinman, 1998). DCs are one of two types of mononuclear phagocytes that populate most tissues, the other being macrophages. The term "phagocyte" derives from the Greek word "phago", meaning "to devour", and reflects the ability of DCs and macrophages to capture exogenous proteins and damaged or dying cells. In

☆ This paper was presented at the 3rd Immunoinformatics and Computational Immunology Workshop (ICIW2012), Orlando, Florida, USA October 7, 2012.

* Corresponding author at: Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, One Gustave L. Levy Place, Box 1498, New York, NY 10029, USA. Tel.: +1 212 659 8535; fax: +1 646 537 8660.
E-mail address: gaurav.pandey@mssm.edu (G. Pandey).

0022-1759/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jim.2012.09.012

contrast to macrophages, whose main role is to scavenge phagocytosed material, DCs sample the local environment to inform the adaptive immune system about peripheral cues. They constantly transport environmental proteins broken down into small peptides termed "antigens" to the lymph node. There, they present self and foreign antigens on MHC-class I- and MHC-class II peptide complexes on the cell surface to resident lymphocytes and produce large amounts of activating cytokines (Guermonprez et al., 2002; Trombetta and Mellman, 2005) to promote the differentiation of antigen-specific effector immune responses (Steinman and Banchereau, 2007). In the case of the presentation of self-antigens, DCs cause the differentiation of antigen-specific T regulatory cells or the depletion of auto-reactive T cells 1(Steinman et al., 2003). MHC-class I and MHC-class II peptides are presented by DCs to induce a CD8+ or CD4+ T cell response respectively. CD8 + T cells are cytotoxic T cells, which specialize in the elimination of infected cells and thus are geared to respond to intracellular pathogens, while CD4+ T cells initiate antibody production of antigen-specific antibodies by B cells to respond to extracellular pathogens. Clearly, DCs play a key role in directing effective immune responses. However, the study of DCs has been hampered due to their rarity within tissues and, until recently, the inability to distinguish DCs from other tissue phagocytes such as macrophages.

Recent data have established that DCs consist of distinct subsets with different abilities to process antigens, respond to environmental stimuli and engage distinct effector lymphocytes (Heath and Carbone, 2009). The DC population can be divided into the following subsets based on ontogeny and function: plasmacytoid DCs (pDCs) and classical DCs (cDCs). These cells arise from different origins in the immune cell lineage and serve specialized immunological functions. pDCs secrete large amounts of the antiviral interferon alpha (IFN-α) cytokine in response to the stimulation of pathogen recognition receptors TLR7 and TLR9 to initiate T cell immunity against viral antigens (Reizis et al., 2011). These cells express low levels of MHC-II and the co-stimulatory cytokines needed to activate T cells in steady state tissue. In contrast, cDCs express high levels of MHC as well as co-stimulatory molecules and are the only hematopoietic cell population with the ability to stimulate naïve T cells in the steady state. Other hematopoietic populations can only stimulate T cells that have already been exposed to antigen, or "memory T cells".

In lymphoid tissue, cDCs consist of two main subsets, namely the CD8+ and CD8– DCs. CD8+ cDCs excel in the cross-presentation of cell-associated antigens and are most potent at stimulating $CD8^+$ T cells to induce a Th1 response (Coombes and Powrie, 2008). This population relies on the cytokine receptor Flt3 and the transcription factors ID2, Batf3, and Irf8 for development. In contrast, CD8– cDCs are most potent at inducing CD4+ T cells to induce a Th2 response (Heath and Carbone, 2009). This population requires Irf4 for their development (Reizis et al., 2011). Recent data established that CD8– DCs are very likely heterogeneous and include at least two main populations that are differentially controlled by Notch2 signaling (Lewis et al., 2011), thus making them very difficult to study.

Owing to the low numbers of DCs in tissues, the difficulty of isolating them from peripheral tissues, and the general expense of these procedures, most DC studies have been limited to the spleen with a limited number of replicates. Through targeted, generally low-throughput, studies, several genes have been identified to be involved in the functioning of DCs and their response to antigens. These genes, several of which are known regulators, include Relb, Irf8, Id2 and Flt3 (Shortman and Heath, 2010). Recent studies have employed high-throughput technologies, such as microarrays, to understand DC biology in vivo. This has greatly accelerated the study of DCs by 1) identifying subset-specific regulators, including Batf3 (Hildner et al., 2008), and most recently, Zbtb46 (Meredith et al., 2012; Satpathy et al., 2012), 2) showing that DC subsets differentially express important surface receptors and regulators (Edwards et al., 2008) and 3) that genes characteristic of the various DC subsets are conserved (Contreras et al., 2010). However, these studies utilize single gene analyses, such as measuring differential expression, to identify genes important for the functioning of DCs (Bar-On et al., 2010; Crozat et al., 2010; Manicassamy et al., 2010; Chevrier et al., 2011). Clearly, such approaches do not reveal the interactions between genes that are equally critical for this problem, as has been done for other immune cell types by systems biology approaches (Amit et al., 2009; Germain et al., 2011; Novershtern et al., 2011; Benichou et al., 2012).

Motivated by the need to build models for DC function that reveal cellular interactions in addition to important genes, we propose a systematic approach that derives detailed modules and networks from large-scale gene expression data sets. For this, we use the WGCNA algorithm (Langfelder and Horvath, 2008) to cluster the relevant portion of the ImmGen gene expression data into groups of tightly co-expressed genes, also known as co-expression modules.[1] We further convert these modules into co-expression networks, and integrate them with transcriptional regulatory data from the Molecular Signature Database (MSigDB) (Subramanian et al., 2005) and protein interaction data from BioGRID (Stark et al., 2011) to generate integrated networks that can help decipher the regulatory structure of these modules.

We use this approach to derive the first set of DC-focused modules and networks (to the best of our knowledge). For this, we build on Miller et al.'s (2012)'s work, where several insights were revealed about DC subsets, specifically cDCs, pDCs and CD8+'s, as well as overall DC functioning. Using their proposed core signatures for these subsets, we conduct an extensive evaluation of our pipeline to identify the most enriched modules that reveal informative integrated networks consisting of many genes and their co-expression and regulatory interactions. A detailed examination of these networks and modules highlights several novel genes, as well as interactions, that may explain the functioning of cDC, pDC and CD8+ cells, and thus add valuable knowledge to DC biology.

In summary, through the example of dendritic cells, we demonstrate how established algorithms and data sources can help generate actionable hypotheses about critical immunological processes, especially involving cell types that are under-represented in data sets.

[1] The terms "cluster" and "module" will be used interchangeably in the rest of this paper.

2. Materials and methods

Fig. 1 shows the different steps of the analysis pipeline and their inputs and outputs. Below, we explain each of the steps of this pipeline and the rationale behind them.

2.1. ImmGen data and modifications

The ImmGen data set (Heng and Painter, 2008) was prepared by sorting multiple replicates of 262 cell populations (1 population = 1 cell type extracted from 1 tissue) from mice and profiling their genome-wide gene expression profiles using the Affymetrix Mouse Gene 1.0 ST array. The raw data were normalized using the RMA algorithm, resulting in a gene expression data matrix spanning 25,194 genes and 853 samples.

For the purpose of this study, we eliminated several samples corresponding to cell populations that could not be classified clearly as either DC or non-DC, reducing the size of the data set to 680 samples. Of these, 56 samples corresponded to DCs (11 pDCs, 45 cDCs and 28 CD8s (subset of cDCs)), and 624 to non-DC cell types (T-cells, B-cells, etc.). Due to this significant imbalance between the two categories, we anticipated the resultant analysis results to be biased in favor of the non-DC cell types (Xiong et al., 2009). Thus, to reduce the potential effect of this bias, we averaged the expression profiles of all the replicates for each of the non-DC cell populations into one sample each, thus reducing their number to 190 samples. Evaluation of the clusters derived from this "compressed" data set demonstrates that this indeed improved our ability to discover more DC-specific modules (Section 3.1).

2.2. Module discovery using WGCNA

The Weighted Correlation Network Analysis (WGCNA) algorithm (Langfelder and Horvath, 2008) and its variants (Zhang and Horvath, 2005) have proven to be very effective for deriving groups of highly co-expressed genes, also referred to as co-expression modules, from large gene expression data sets (Miller et al., 2010; Voineagu et al., 2011). WGCNA begins with a matrix of the absolute value of the Pearson correlation coefficients between all gene pairs, and converts this matrix into an adjacency matrix using a power function $f(x) = x^{\beta}$. The parameter β of the power function is determined in such a way that the resulting adjacency matrix (i.e., the weighted co-expression network) is approximately scale-free, a widely accepted property of biological networks. To measure how well a network satisfies a scale-free topology, we use the fitting index (Zhang and Horvath, 2005), i.e., the model fitting index R^2 of the linear model that regresses $\log(p(k))$ on $\log(k)$, where k is connectivity and $p(k)$ is the frequency distribution of connectivity. The fitting index of a perfect scale-free network is 1. In our analysis, we selected the smallest β that leads to the highest R^2 (an approximately scale-free network), and this value turned out to be 7 ($R^2 = 0.769$) and 6 ($R^2 = 0.723$) for the compressed and original DC expression data sets respectively.

To explore the modular structures of the co-expression network, the adjacency matrix is further transformed into a

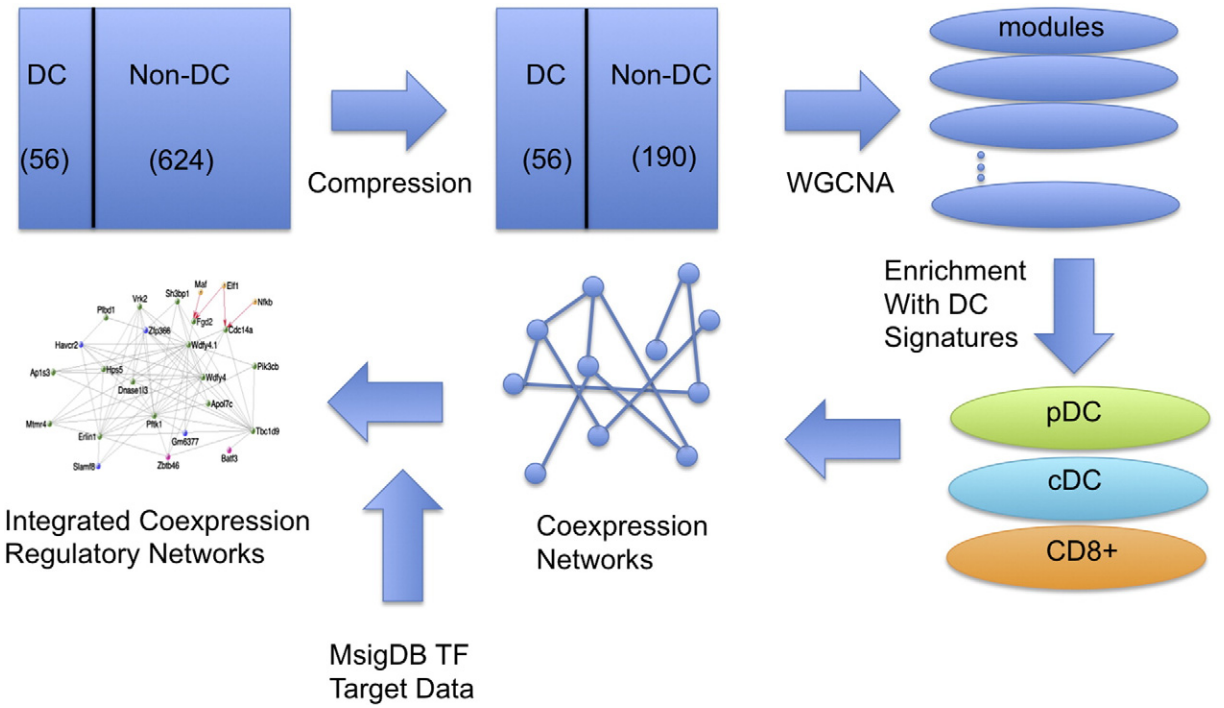

Fig. 1. Our analysis pipeline.

matrix of topological overlap measures (TOM) (Zhang and Horvath, 2005). The TOM score between genes i and j is defined as:

$$TOM(i,j) = \frac{l_{ij} + a_{ij}}{\min\left(k_i, k_j\right) + 1 - a_{ij}}$$

Here, a_{ij} is the adjacency score between i and j (calculated as described above), $k_i = \sum_u a_{iu}$ (sum of edge weights) and $l_{ij} = \sum_u a_{iu} a_{uj}$ (sum of the product of edge weights involving all common neighbors). Thus, TOM measures the strength of the association between two genes in a graph based on the ratio of the similarity of their common neighborhood (numerator) to the smaller of the individual neighborhoods of the two nodes (denominator). It also gives higher weight to genes that are already strongly associated (high a_{ij}) in the input adjacency matrix.

WGCNA finds the final modules of highly co-expressed genes using average linkage hierarchical clustering to group genes based on this topological overlap matrix. It then dynamically groups closely related modules in the resultant dendrogram using a dynamic cut-tree algorithm (Langfelder et al., 2008). Previous studies (Ravasz et al., 2002; Zhang and Horvath, 2005) have shown that WGCNA leads to more cohesive and biologically meaningful modules than other clusterings based on Pearson correlation due to its robustness to noise and better ability to capture indirect associations between genes. This motivated the use of this module discovery algorithm in our study.

We applied the WGCNA algorithm (Langfelder and Horvath, 2008) to our compressed data set to identify such modules of co-expressed genes. The algorithm was applied with a block size of 2000 and merge cut height of 0.[2] We also generated several sets of modules by setting the minModuleSize parameter to 10, 20 and 30, since this parameter appeared to affect the size and number of the modules substantially. For comparison, we also derived modules from the original uncompressed data set using WGCNA with the same parameter settings, as well as using hierarchical clustering with a very similar methodology.

2.3. Functional analysis of modules

To understand the direct relationship of these modules with the immune system and the functioning of dendritic cells, we next investigated their significance in the context of Miller et al.'s (2012) results about DC biology. Here, through a principal component analysis of the ImmGen data set, the authors showed that DC segregated into distinct populations, specifically cDCs, pDCs and CD8+ cells. By comparing these subsets against their closest cell type in the principal component map, the authors identified core signatures that consist of genes that are significantly differentially expressed between the two classes for each of these subsets. Using the Fisher exact test, Bonferroni correction (correction factor = number of modules discovered) and the set of genes clustered as the background set, we evaluated how many of the modules found using WGCNA (and hierarchical clustering) were enriched for these core signatures (corrected p-value < 0.05), and used this number to identify the most informative set of modules to study further. Note that although we are using the core signatures derived from the same data set for this evaluation, the primary purpose of the modules is to uncover genes and interactions critical to DC function. The signatures are used only as a guide towards achieving this larger goal.

In the most informative set of modules, we further identified one module for each DC subset for further study, and evaluated the specificity of these modules to the corresponding DC subset. For this, we calculated the median of the individual Student's t-test p-values of the genes constituting each module as a measure of its differential expression between the DC subset assigned and all the other samples in our compressed data set. The sample labels (1 for DC subset and 0 for the others) were then randomly permuted 1000 times, and the differential expression measure of the module recomputed. The final differential expression p-value of the module was then assigned to be the fraction of random permutations in which the measure was lower than the value for the original sample labels. Once we ensured that these modules were indeed specific to the corresponding DC subset, we chose them as the representative modules to be further analyzed in terms of their co-expression and regulatory structure. Note that although we focus only on these modules in the rest of the paper, similar analyses can be performed for the other enriched modules as well. Those analyses are out of the scope of this paper, but we intend to perform them in future work.

2.4. Network and regulatory analysis

As the first step towards explaining the modules selected above, we conducted a regulatory analysis of these modules by identifying transcription factors (TFs) that are predicted to regulate genes in these modules. For this, we identified the TFs in the Molecular Signature Database (MSigDB) (Mootha et al., 2003; Subramanian et al., 2005) whose known target genes had a significant overlap (Fisher's exact test, Benjamini–Hochberg correction) with the genes in the module being considered. TFs with p-values less than 0.05 were considered to be regulating the expression of their target genes in the module, which are also included in the MSigDB search result. The representation of the connections between these TFs and their targets in the modules by directed edges produced regulatory networks for these modules, which provided the first indications for genes and regulators important for the corresponding DC subset.

These networks, although very useful, do not cover too many genes in the modules, owing to the general lack of regulatory information in public databases. Thus, to obtain a more complete view of the interactions underlying DC function, we converted the modules into co-expression network. Constituting each module discovered by WGCNA is a gene-by-gene matrix containing the TOM score for each pair of genes included in the module. This value indicates the strength of the co-expression link between the two genes in the data set analyzed, and thus can be used as an indication of how these genes are functioning together in the

[2] See Langfelder and Horvath (2008) for the details of these parameters of WGCNA and its R implementation. The default values in the implementation were used for the other parameters not mentioned here.

corresponding DC subset. Treating these TOM values as weights of co-expression edges, we combined the corresponding set of co-expression edges with the TF-target links identified earlier to create an integrated network for each module. Where available, we also incorporated protein–protein interactions obtained from BioGRID (Stark et al., 2011) into the integrated networks. The resultant networks were studied to identify genes and interactions that are likely to play an important role in defining the function of the corresponding DC subset, and thus would be promising targets for experimental validation and further study.

3. Results

This section details the results of the evaluation of various components of our analysis pipeline, and illustrates how our approach can be used to discover potentially useful hypothesis about biological processes.

3.1. Evaluation of choices for the pipeline

Our analysis pipeline involved several choices that can influence the quality of modules and networks derived from the complex ImmGen gene expression data set that spans thousands of genes and hundreds of immune cell types. Two of the most critical choices were:

i. Whether to use the original data set, that contained a much larger number of non-DC samples as compared with DCs or its compressed version, where the fraction of DC samples is increased while retaining the essential data for the other cell types?
ii. Which clustering algorithm – WGCNA or the more commonly used hierarchical clustering – to use for discovering clusters?

To evaluate these choices, we derived clusters using WGCNA and hierarchical clustering, both from the original and the compressed data set. The same R implementation of WGCNA was used for hierarchical clustering as well, the only difference being that Pearson's correlation coefficient was used as the similarity measure instead of TOM. We set the value of the minModuleSize parameter to 10, 20 and 30 to examine its influence on the number, size and quality of the modules obtained. The resultant sets of modules were then evaluated in terms of how many DC-enriched modules they include, as described in Section 2.2.

Tables 1 and 2 show the results of this evaluation for WGCNA and hierarchical clustering respectively. In both the tables, it can be seen that, at the same value of minModuleSize, the compressed data set produces several more modules enriched with the core signature of at least one DC subset (last column), thus demonstrating the utility of the compression step of our pipeline. Furthermore, comparing the two tables, WGCNA discovers several more DC-enriched modules than hierarchical clustering, almost regardless of the value of minModuleSize. For example, WGCNA and hierarchical clustering discover 152 and 164 modules from the compressed data set at minModuleSize = 20 and 10 respectively. Despite the slightly smaller number of modules, 8 of the WGCNA modules are DC-enriched, as against 6 of the hierarchical clustering modules. This better ability of WGCNA to discover more meaningful modules is due to its use of the TOM measure, which is better able to resist the noise in large gene expression data sets as compared with direct similarity measures like Pearson's correlation coefficient (Zhang and Horvath, 2005). Finally, setting the value of minModuleSize to 10 produces the most DC-enriched modules for both data sets and clustering algorithms, since this value produces the smallest modules that are more likely to capture specific biological processes (Pandey et al., 2009).

Table 1
Statistics about the number of all the modules and the DC-enriched ones discovered by WGCNA for different choices of the data set and values of minModuleSize.

			# Enriched modules			
Data	minModuleSize	# Modules	cDC	pDC	CD8	Any
Original	10	166	5	2	1	7
	20	120	5	2	1	7
	30	98	5	1	1	6
Compressed	10	218	8	1	2	10
	20	152	7	1	1	8
	30	104	4	1	1	5

Table 2
Statistics about the number of all the modules and the DC-enriched ones discovered by hierarchical clustering for different choices of the data set and values of minModuleSize.

			# Enriched modules			
Data	minModuleSize	# Modules	cDC	pDC	CD8	Any
Original	10	147	3	1	1	4
	20	64	2	1	1	3
	30	40	1	1	0	2
Compressed	10	164	4	2	1	6
	20	67	3	1	1	4
	30	48	3	1	1	4

Previous studies using WGCNA for discovering co-expression modules pre-selected a subset of genes to reduce the adverse effects of including genes with invariable and noisy expression profiles (Miller et al., 2008; de Jong et al., 2010; Ye et al., 2012). To test this approach for our study, we selected several subsets of genes whose expression profiles showed the highest variance in the compressed data set and discovered WGCNA modules from the resultant data sets. Table 3 shows the results of the DC enrichment evaluation on the modules discovered. As the number of selected genes increased, the number of modules discovered also increased naturally, but more interestingly, the number of DC-enriched modules (last column) also increased, with the full data set producing the most such modules.

Based on the results of these evaluations, we selected the modules discovered from the compressed data set (all genes) using WGCNA (minModuleSize = 10) to study DC function further.

3.2. Characteristics of DC-enriched modules

Table 1 shows that our pipeline is able to discover several modules enriched for Miller et al.'s (2012) DC subset-specific core signatures. To understand the functioning of these DC subsets, we focused on the most enriched (lowest p-value)

Table 3
Comparison of the number of DC-enriched modules discovered using different number of pre-selected genes from the data set.

# Genes	# Modules	# Enriched modules			
		cDC	pDC	CD8+	Any
25,194 (All)	218	8	1	2	10
15,000	173	7	1	1	8
10,000	117	7	1	2	9
8000	101	6	1	1	7
6000	90	5	1	1	6
4000	65	3	0	1	3
2000	29	0	0	0	0

modules for each of these subsets. Table 4 lists various statistics for the three representative modules so identified, and Supplementary Table 1 contains all the member genes for all these modules.

It is evident from Table 4 that these modules are very highly enriched for the core signatures of the corresponding subsets, as indicated by the very low enrichment p-values and high fold changes. Furthermore, the differential p-values of all these modules are very low ($<10^{-3}$). This means that the majority of the genes in these modules are significantly differentially expressed between the corresponding DC subset and other samples, thus showing that they are very specific to the corresponding DC subset. This is significant, since no information about the original DC subset of the samples was used to derive these modules from the compressed data set. We also tested these modules for enrichment with Gene Ontology terms (Boyle et al., 2004), but were unable to find any informative functions for them. This is most likely due to the lack of annotations related to immunology in general, and dendritic cell function in particular, in biological knowledge bases. We hope that our detailed study of the genes and interactions constituting these modules, discussed in the next section, and other such studies will help fill this gap in knowledge about immunological processes.

3.3. Network and regulatory analysis of DC-enriched modules

The statistical analysis of the representative modules (Table 4) indicated their utility for studying DC biology. However, this analysis does not provide much information about the interactions between the constituent genes and with other genes, such as transcription factors (TFs), that lead to the corresponding functions being performed. Thus, we attempted to explain the functioning of these modules in terms of these interactions.

First, we identified the TFs regulating the expression of the genes in each module by searching the Molecular Signature Database (MSigDB). We identified 3 and 4 TFs for the cDC and CD8+ representative modules using a p-value threshold of 0.05, but had to raise the threshold to 0.1 to identify 2 significant TFs for the pDC module due to its relatively large size. Fig. 2 shows the directed networks representing the links

Table 4
Details of modules found to be most enriched for DC subsets.

DC subset	Enrichment p-value	Enrichment fold change	# Subset genes	Differential p-value	Size
cDC	0	42.39	16	$<10^{-3}$	70
pDC	0	33.39	54	$<10^{-3}$	428
CD8+	0	119.46	14	$<10^{-3}$	94

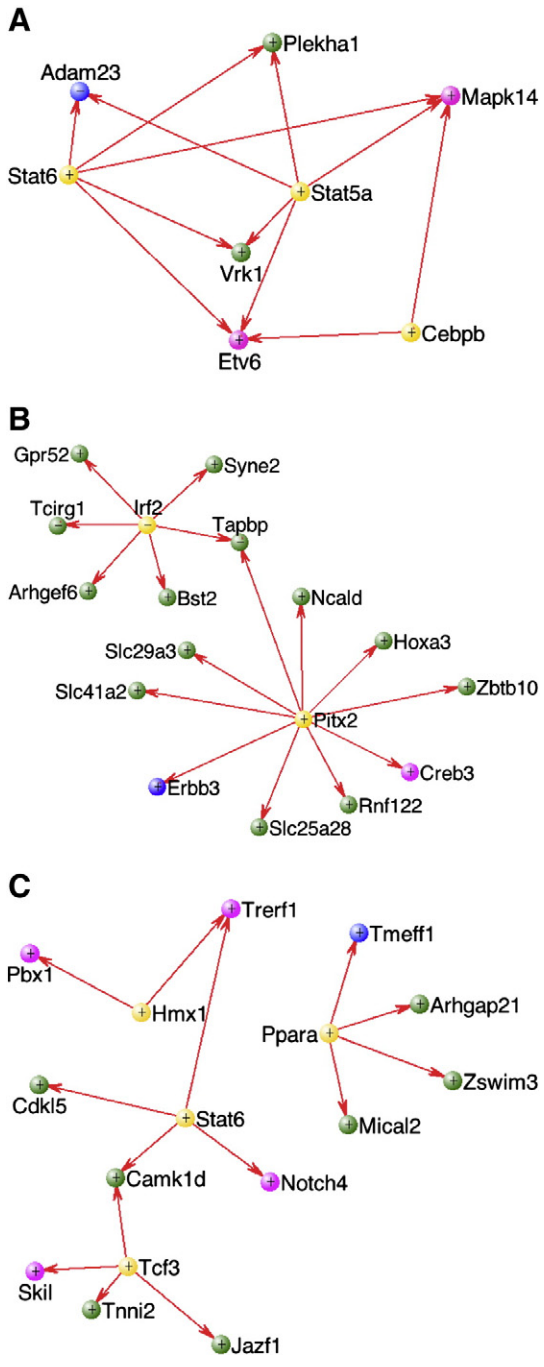

Fig. 2. Regulatory (TF-target) networks for the (A) cDC (B) pDC and (C) CD8+ representative modules discovered by our pipeline. Genes colored in blue are found within the corresponding core signature, transcription factors from MSigDB are colored yellow, predicted regulators not found in MSigDB searches are colored pink and all the other genes are shown as green. Red directed edges denote TF-target links.

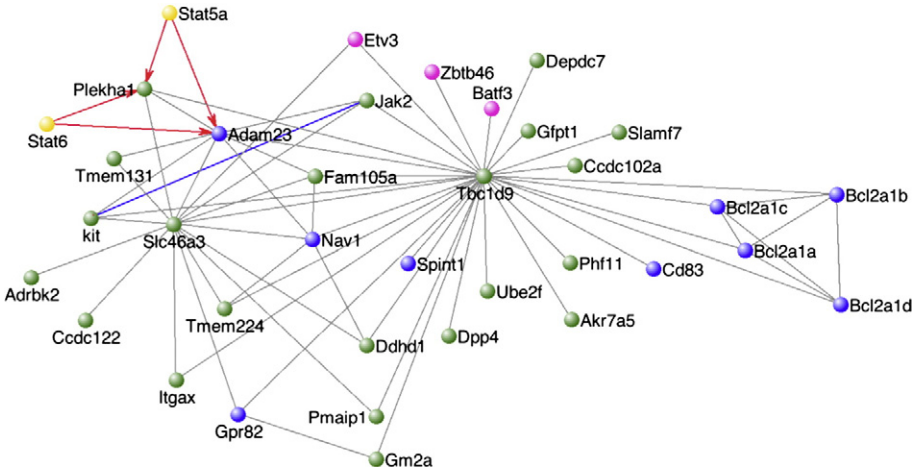

Fig. 3. Integrated (co-expression, TF-target and protein–protein) network derived from the representative cDC module. Genes colored in blue are found within the corresponding core signature, transcription factors from MSigDB are colored yellow, predicted regulators not found in MSigDB searches are colored pink are and all the other genes are shown as green. Undirected gray edges denote co-expression links between genes, red directed edges denote TF-target links derived from MSigDB, and blue edges denote protein–protein interactions. For visual clarity, only co-expression edges with TOM scores higher than 0.1 are included. The complete network can be found in Supplementary Table 1.

between these TFs and their targets in these modules. To address the lack of sufficient TF-target information in public databases, and to identify potentially non-regulatory interactions between genes in these module, we constructed networks consisting of co-expression links between the constituent genes. Treating the genes as nodes, the TOM score for two genes as the weight of the edge between them, and the TOM matrix as the overall adjacency matrix, each module can be viewed as a co-expression network. This network can provide very useful insights into the functioning of a module, since coherent expression of two genes is a strong indicator of them performing the same or similar functions. We also incorporated the MSigDB-derived TF-target links, and protein–protein interactions from BioGRID into these networks to obtain integrated networks for the representative modules, details of which can be found in Supplementary Table 1. Figs. 3–5 show these integrated networks corresponding to the strongest co-expression links within the cDC, pDC and CD8+ modules, where the links with TOM scores higher than 0.1, 0.4 and 0.2 were included. This selection was done to enhance visual clarity, as well as to focus on the strongest co-expression links, which are the most likely to indicate functional relevance. Some genes that are predicted to be regulators in the mouse genome (Novershtern et al., 2011), but may not be included in MSigDB, are highlighted (colored pink) in these networks. Furthermore, all these networks

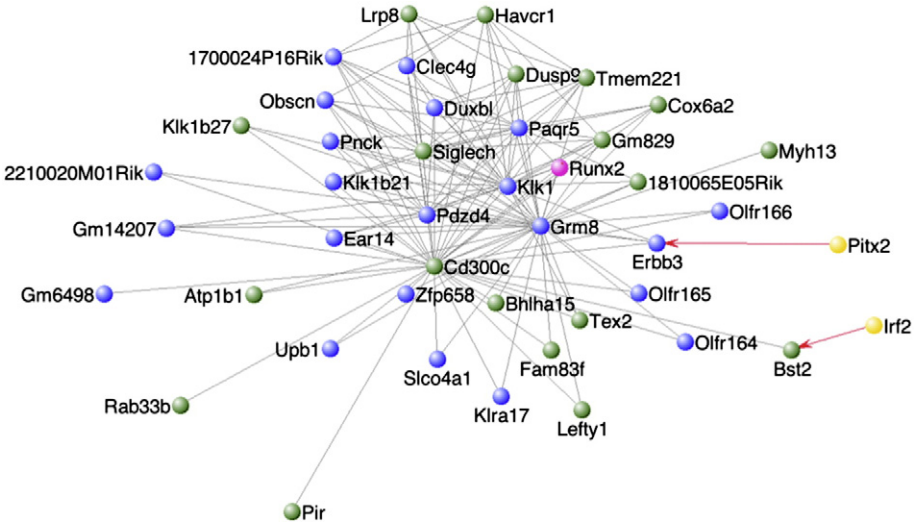

Fig. 4. Integrated (co-expression, TF-target and protein–protein) network derived from the representative pDC module. Genes colored in blue are found within the corresponding core signature, transcription factors from MSigDB are colored yellow, predicted regulators not found in MSigDB searches are colored pink are and all the other genes are shown as green. Undirected gray edges denote co-expression links between genes and red directed edges denote TF-target links derived from MSigDB. For visual clarity, only co-expression edges with TOM scores higher than 0.4 are included. The complete network can be found in Supplementary Table 1.

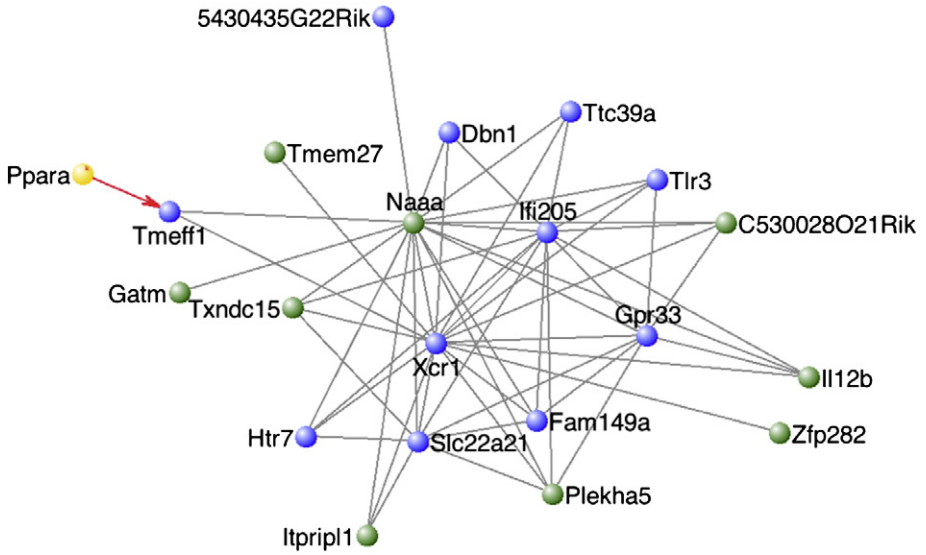

Fig. 5. Integrated (co-expression, TF-target and protein–protein) network derived from the representative CD8+ module. Genes colored in blue are found within the corresponding core signature, transcription factors from MSigDB are colored yellow, predicted regulators not found in MSigDB searches are colored pink are and all the other genes are shown as green. Undirected gray edges denote co-expression links between genes and red directed edges denote TF-target links derived from MSigDB. For visual clarity, only co-expression edges with TOM scores higher than 0.2 are included. The complete network can be found in Supplementary Table 1.

include many genes (colored in blue) included in Miller et al.'s (2012) core signatures for the DC subsets, thus providing some visual validation of the results in Table 4. Examining the networks in detail provides several interesting pieces of information about DC biology.

cDCs have a superior ability in comparison with all immune cells to present antigen on both MHC (major histocompatibility complex) Class I and MHC Class II to induce a CD8+ and a CD4+ T cell response respectively. They are the only population able to present antigen to naïve T cells to mount an immune response and this is made possible by their production of key T cell skewing co-stimulatory cytokines, in addition to their antigen presentation capabilities. However, the mechanisms underlying both these cDC abilities are still unclear. Intriguingly, the cDC TF network (Fig. 2(a)) identified the signal transducers and activators of transcription Stat5a and Stat6 as regulators of cDC. Furthermore, these genes, along with Jak2, interact with the metalloproteinase Adam23 in the integrated cDC network (Fig. 3). Jak2-deficient mice have decreased numbers of cDCs and the remaining cDCs express lower levels of cDC maturation markers, which include genes CD80 and CD83 found in our representative module for this subset. The loss of Jak2 signaling along with either Stat5a or Stat6 leads to an impairment of the production of crucial co-stimulatory cytokines TNFα and IL-12, thus leading to the loss of cDC function (Zhong et al., 2010). The interaction of these regulators with members of this module, captured by the integrated network, may provide explanations for this observation, and insights into the regulatory functions of cDCs. This network also includes known cDC regulators Batf3 and Zbtb46 (Hildner et al., 2008; Meredith et al., 2012; Satpathy et al., 2012), as well as the phenotypical marker Cd11c (Itgax) for this subset (Hashimoto et al., 2011). Batf3 has been shown to be expressed in all cDC populations, and the loss of this gene in murine knockout models selectively prevents the development of CD8+ DC (Hashimoto et al., 2011). More such candidate genes and mechanisms are expected to be discovered by a detailed examination of the proposed cDC module and networks.

pDCs produce paramount levels of IFN-a (Interferon-alpha) in response to stimuli through TLRs (Toll-like receptors) 7 and 9 (Reizis et al., 2011). They have additionally been shown to induce both CD4+ and CD8+ T cells and, conversely, tolerance. These seemingly contradictory functions are likely due to a wide array of surface markers that allow the pDCs to respond uniquely to various stimuli and environments (Reizis et al., 2011). In accordance with this theory, the pDC module and network (Fig. 4) contain multiple genes known to be involved in the various proposed pDC functions, including IFN-a inducing genes Klra17 (Ly49Q) (Reizis et al., 2011) and CD300c (Ju et al., 2008), as well as the IFN-attenuating receptor Siglec-H (Reizis et al., 2011). pDCs also express the CD4+ T cell stimulatory molecule Havcr1 (Rodriguez-Manzanet et al., 2009), as well as high levels of Lag3 shown to be important in the peripheral CD8+ T cell tolerance induction (Lucas et al., 2011). The interactions of the predicted regulator Runx2, which is highly specific to pDCs (Reizis et al., 2011), can help identify its as yet undefined role in this DC subset. Further examination of this module and its corresponding network can highlight interactions between the genes discussed here, as well as the other constituent genes with no known immune function that can facilitate the better understanding of pDCs.

CD8+ DCs are a subset of cDCs specialized in the uptake and cross-presentation of antigen from apoptotic cells on MHC-I in order to mount a CD8+ T cell response. In accordance with this function, they produce the CD8+ T cell inducing cytokine IL-12b (Rosenblum et al., 2010), as well as high levels of TLR3, which binds to double stranded RNA found in retroviruses that replicate intra-cellularly (Shortman and Heath, 2010). The CD8+ integrated network (Fig. 5) includes

the corresponding genes Tlr3 and Il12b (Rosenblum et al., 2010), thus providing evidence for the network's validity and utility. Furthermore, this module includes Xcr1, a chemokine shown to control CD8 + T cell effector differentiation (Dorner et al., 2009). Pbx1, a predicted regulator of this module (Fig. 2(c)) has been shown to be important in induction of the T cell suppressing cytokine IL-10 production in response to apoptotic debris. This may contribute to the tolerogenic function attributed to DC in the steady state (Shortman and Heath, 2010), thus indicating a role for Pbx1 in CD8 + function. Finally, in addition to the statistical evidence discussed earlier, these examples reveal that the CD8 + DC module is enriched for genes involved in the cardinal function of DCs, antigen processing, as well as transcripts involved in lysosome function, an organelle essential for antigen processing, further validating our approach.

These specific instances indicate the utility of these co-expression and regulatory networks for identifying new genes and regulators that may control DC function and specialization in vivo. In addition, they provide a network or system context for understanding how genes, both known and novel, affect DC function by interacting with or regulating other genes.

4. Conclusions and discussion

In this paper, we reported the results of our module and network analysis of the ImmGen gene expression data set, with the goal of extracting novel insights about the functioning of dendritic cells (DCs), as well as their subsets, namely cDC, pDC and CD8 + cells. After compressing the non-DC component of this data set, we applied the WGCNA algorithm to it to identify modules (clusters) of co-expressed genes that are expected to be involved in DC-related processes. In particular, several of these modules are found to be significantly enriched for the core signatures of the cDC, pDC and CD8 + subsets, and are also significantly differentially expressed with respect to these subsets. We conducted further network analysis of these modules by viewing them as co-expression networks and integrating them with transcription factor-target links derived from the MSigDB database. This analysis highlighted several genes whose position and interactions in the networks indicated their importance for the functioning of the corresponding DC subset. Overall, having such a network view for how different genes and regulators interact with each other within a modular context can provide novel insights into the mechanisms underlying immunological processes in general, and DC function and differentiation in particular.

Although our study did not include experimental validation, we believe that the proposed networks and modules include several genes critical to DC function, especially the ones for which literature evidence was presented. The proposed networks can be analyzed in terms of their structure to prioritize validation targets, some work on which is ongoing in our group. The functional information obtained from such analysis and validation can help address the general lack of immunological information in public databases, such as Gene Ontology, especially for the mouse genome. Another effort that can help in this direction is using our approach, which is not specific to mouse DCs, to study other immune cell types from other organisms as well, especially those under-represented in studies and data sets. This will help identify organism-specific factors influencing their immune systems.

Finally, although our approach was rigorous, it was limited to the ImmGen gene expression data set. The approach, as well as the results obtained from it, can become a lot more powerful, if other omics data, such as those from next-generation sequencing, proteomics and metabolomic technologies, are also incorporated into the pipeline. The use of next-generation sequencing technologies, such as RNA-Seq (Wang et al., 2009), can help expand the coverage of transcripts as compared with microarrays. Such integration has been very successful in numerous other areas (Hawkins et al., 2010; Kasarskis et al., 2011). The resultant computational and systems biology approaches will be immensely useful for understanding the components of the immune system, as well as the interactions between them. We expect a rapid growth in this area in the time to come.

Supplementary data related to this article can be found online at http://dx.doi.org/10.1016/j.jim.2012.09.012.

Acknowledgments

We gratefully acknowledge the financial and technical support of the Institute for Genomics and Multiscale Biology at Mount Sinai.

References

Amit, I., Garber, M., Chevrier, N., Leite, A.P., Donner, Y., Eisenhaure, T., Guttman, M., Grenier, J.K., Li, W., Zuk, O., Schubert, L.A., Birditt, B., Shay, T., Goren, A., Zhang, X., Smith, Z., Deering, R., McDonald, R.C., Cabili, M., Bernstein, B.E., Rinn, J.L., Meissner, A., Root, D.E., Hacohen, N., Regev, A., 2009. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. Science 326, 257.

Banchereau, J., Steinman, R.M., 1998. Dendritic cells and the control of immunity. Nature 392, 245.

Bar-On, L., Birnberg, T., Lewis, K.L., Edelson, B.T., Bruder, D., Hildner, K., Buer, J., Murphy, K.M., Reizis, B., Jung, S., 2010. CX3CR1 + CD8alpha + dendritic cells are a steady-state population related to plasmacytoid dendritic cells. Proc. Natl. Acad. Sci. U. S. A. 107, 14745.

Benichou, J., Ben-Hamo, R., Louzoun, Y., Efroni, S., 2012. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. Immunology 135, 183.

Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M., Sherlock, G., 2004. GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 20, 3710.

Chevrier, N., Mertins, P., Artyomov, M.N., Shalek, A.K., Iannacone, M., Ciaccio, M.F., Gat-Viks, I., Tonti, E., DeGrace, M.M., Clauser, K.R., Garber, M., Eisenhaure, T.M., Yosef, N., Robinson, J., Sutton, A., Andersen, M.S., Root, D.E., von Andrian, U., Jones, R.B., Park, H., Carr, S.A., Regev, A., Amit, I., Hacohen, N., 2011. Systematic discovery of TLR signaling components delineates viral-sensing circuits. Cell 147, 853.

Contreras, V., Urien, C., Guiton, R., Alexandre, Y., Vu Manh, T.P., Andrieu, T., Crozat, K., Jouneau, L., Bertho, N., Epardaud, M., Hope, J., Savina, A., Amigorena, S., Bonneau, M., Dalod, M., Schwartz-Cornil, I., 2010. Existence of CD8alpha-like dendritic cells with a conserved functional specialization and a common molecular signature in distant mammalian species. J. Immunol. 185, 3313.

Coombes, J.L., Powrie, F., 2008. Dendritic cells in intestinal immune regulation. Nat. Rev. Immunol. 8, 435.

Crozat, K., Guiton, R., Contreras, V., Feuillet, V., Dutertre, C.A., Ventre, E., Vu Manh, T.P., Baranek, T., Storset, A.K., Marvel, J., Boudinot, P., Hosmalin, A., Schwartz-Cornil, I., Dalod, M., 2010. The XC chemokine receptor 1 is a conserved selective marker of mammalian cells homologous to mouse CD8alpha + dendritic cells. J. Exp. Med. 207, 1283.

de Jong, S., Fuller, T.F., Janson, E., Strengman, E., Horvath, S., Kas, M.J., Ophoff, R.A., 2010. Gene expression profiling in C57BL/6J and A/J mouse inbred strains reveals gene networks specific for brain regions independent of genetic background. BMC Genomics 11, 20.

Dorner, B.G., Dorner, M.B., Zhou, X., Opitz, C., Mora, A., Guttler, S., Hutloff, A., Mages, H.W., Ranke, K., Schaefer, M., Jack, R.S., Henn, V., Kroczek, R.A., 2009. Selective expression of the chemokine receptor XCR1 on cross-presenting dendritic cells determines cooperation with CD8+ T cells. Immunity 31, 823.

Edwards, A.G., Weale, A.R., Denny, A.J., Edwards, K.J., Helps, C.R., Lear, P.A., Bailey, M., 2008. Antigen receptor V-segment usage in mucosal T cells. Immunology 123, 181.

Gardy, J.L., Lynn, D.J., Brinkman, F.S., Hancock, R.E., 2009. Enabling a systems biology approach to immunology: focus on innate immunity. Trends Immunol. 30, 249.

Germain, R.N., Meier-Schellersheim, M., Nita-Lazar, A., Fraser, I.D., 2011. Systems biology in immunology: a computational modeling perspective. Annu. Rev. Immunol. 29, 527.

Guermonprez, P., Valladeau, J., Zitvogel, L., Thery, C., Amigorena, S., 2002. Antigen presentation and T cell stimulation by dendritic cells. Annu. Rev. Immunol. 20, 621.

Hashimoto, D., Miller, J., Merad, M., 2011. Dendritic cell and macrophage heterogeneity in vivo. Immunity 35, 323.

Hawkins, R.D., Hon, G.C., Ren, B., 2010. Next-generation genomics: an integrative approach. Nat. Rev. Genet. 11, 476.

Heath, W.R., Carbone, F.R., 2009. Dendritic cell subsets in primary and secondary T cell responses at body surfaces. Nat. Immunol. 10, 1237.

Heng, T.S., Painter, M.W., 2008. The Immunological Genome Project: networks of gene expression in immune cells. Nat. Immunol. 9, 1091.

Hildner, K., Edelson, B.T., Purtha, W.E., Diamond, M., Matsushita, H., Kohyama, M., Calderon, B., Schraml, B.U., Unanue, E.R., Diamond, M.S., Schreiber, R.D., Murphy, T.L., Murphy, K.M., 2008. Batf3 deficiency reveals a critical role for CD8alpha+ dendritic cells in cytotoxic T cell immunity. Science 322, 1097.

Ju, X., Zenke, M., Hart, D.N., Clark, G.J., 2008. CD300a/c regulate type I interferon and TNF-alpha secretion by human plasmacytoid dendritic cells stimulated with TLR7 and TLR9 ligands. Blood 112, 1184.

Kasarskis, A., Yang, X., Schadt, E., 2011. Integrative genomics strategies to elucidate the complexity of drug response. Pharmacogenomics 12, 1695.

Langfelder, P., Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559.

Langfelder, P., Zhang, B., Horvath, S., 2008. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut library for R. Bioinformatics 24 (5), 719.

Lewis, K.L., Caton, M.L., Bogunovic, M., Greter, M., Grajkowska, L.T., Ng, D., Klinakis, A., Charo, I.F., Jung, S., Gommerman, J.L., Ivanov, I.I., Liu, K., Merad, M., Reizis, B., 2011. Notch2 receptor signaling controls functional differentiation of dendritic cells in the spleen and intestine. Immunity 35, 780.

Lucas, C.L., Workman, C.J., Beyaz, S., LoCascio, S., Zhao, G., Vignali, D.A., Sykes, M., 2011. LAG-3, TGF-beta, and cell-intrinsic PD-1 inhibitory pathways contribute to CD8 but not CD4 T-cell tolerance induced by allogeneic BMT with anti-CD40L. Blood 117, 5532.

Manicassamy, S., Reizis, B., Ravindran, R., Nakaya, H., Salazar-Gonzalez, R.M., Wang, Y.C., Pulendran, B., 2010. Activation of beta-catenin in dendritic cells regulates immunity versus tolerance in the intestine. Science 329, 849.

Meredith, M.M., Liu, K., Darrasse-Jeze, G., Kamphorst, A.O., Schreiber, H.A., Guermonprez, P., Idoyaga, J., Cheong, C., Yao, K.H., Niec, R.E., Nussenzweig, M.C., 2012. Expression of the zinc finger transcription factor zDC (Zbtb46, Btbd4) defines the classical dendritic cell lineage. J. Exp. Med. 209, 1153.

Miller, J.A., Oldham, M.C., Geschwind, D.H., 2008. A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging. J. Neurosci. 28, 1410.

Miller, J.A., Horvath, S., Geschwind, D.H., 2010. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. Proc. Natl. Acad. Sci. U. S. A. 107, 12698.

Miller, J., Brown, B., Shay, T., Gautier, E., Jojic, V., Cohain, A., Pandey, G., Leboeuf, M., Elpek, K.G., Helft, J., Hashimoto, D., Chow, A., Price, J., Greter, M., Bogunovic, M., Bellemare-Pelletier, A., Frenette, P.S., Randolph, G.J., Turley, S.J., Merad, M., Consortium, I.G., 2012. Deciphering the transcriptional network of the DC lineage. Nat. Immunol. 13 (9), 888 (Sep).

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D., Groop, L.C., 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. 34, 267.

Novershtern, N., Subramanian, A., Lawton, L.N., Mak, R.H., Haining, W.N., McConkey, M.E., Habib, N., Yosef, N., Chang, C.Y., Shay, T., Frampton, G.M., Drake, A.C., Leskov, I., Nilsson, B., Preffer, F., Dombkowski, D., Evans, J.W., Liefeld, T., Smutko, J.S., Chen, J., Friedman, N., Young, R.A., Golub, T.R., Regev, A., Ebert, B.L., 2011. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell 144, 296.

Pandey, G., Atluri, G., Steinbach, M., Myers, C.L., Kumar, V., 2009. An association analysis approach to biclustering. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Paris, France, p. 677.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Barabasi, A.L., 2002. Hierarchical organization of modularity in metabolic networks. Science 297, 1551.

Reizis, B., Bunin, A., Ghosh, H.S., Lewis, K.L., Sisirak, V., 2011. Plasmacytoid dendritic cells: recent progress and open questions. Annu. Rev. Immunol. 29, 163.

Rodriguez-Manzanet, R., DeKruyff, R., Kuchroo, V.K., Umetsu, D.T., 2009. The costimulatory role of TIM molecules. Immunol. Rev. 229, 259.

Rosenblum, J.M., Shimoda, N., Schenk, A.D., Zhang, H., Kish, D.D., Keslar, K., Farber, J.M., Fairchild, R.L., 2010. CXC chemokine ligand (CXCL) 9 and CXCL10 are antagonistic costimulation molecules during the priming of alloreactive T cell effectors. J. Immunol. 184, 3450.

Satpathy, A.T., Wumesh, K.C., Albring, J.C., Edelson, B.T., Kretzer, N.M., Bhattacharya, D., Murphy, T.L., Murphy, K.M., 2012. Zbtb46 expression distinguishes classical dendritic cells and their committed progenitors from other immune lineages. J. Exp. Med. 209, 1135.

Shortman, K., Heath, W.R., 2010. The CD8+ dendritic cell subset. Immunol. Rev. 234, 18.

Stark, C., Breitkreutz, B.J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X., Reguly, T., Rust, J.M., Winter, A., Dolinski, K., Tyers, M., 2011. The BioGRID Interaction Database: 2011 update. Nucleic Acids Res. 39, D698.

Steinman, R.M., Banchereau, J., 2007. Taking dendritic cells into medicine. Nature 449, 419.

Steinman, R.M., Hawiger, D., Nussenzweig, M.C., 2003. Tolerogenic dendritic cells. Annu. Rev. Immunol. 21, 685.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 102, 15545.

Trombetta, E.S., Mellman, I., 2005. Cell biology of antigen processing in vitro and *in vivo*. Annu. Rev. Immunol. 23, 975.

Voineagu, I., Wang, X., Johnston, P., Lowe, J.K., Tian, Y., Horvath, S., Mill, J., Cantor, R.M., Blencowe, B.J., Geschwind, D.H., 2011. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature 474, 380.

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57.

Xiong, H., Wu, J.J., Chen, J., 2009. K-means clustering versus validation measures: a data-distribution perspective. IEEE Trans. Syst. Man Cybern. B Cybern. 39, 318.

Ye, H., Yu, C.H., Li, L., Xu, C.F., Zhang, X.Q., Li, Y.M., 2012. Meta-analysis of human colorectal cancer transcriptome. Int. J. Colorectal Dis. 27 (8), 1125 (Aug).

Zhang, B., Horvath, S., 2005. A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol. 4 (Article17).

Zhong, J., Yang, P., Muta, K., Dong, R., Marrero, M., Gong, F., Wang, C.Y., 2010. Loss of Jak2 selectively suppresses DC-mediated innate immune response and protects mice from lethal dose of LPS-induced septic shock. PLoS One 5, e9593.