Published online 17 October 2014 *Nucleic Acids Research, 2015, Vol. 43, Database issue* ***D1133–D1139***
doi: 10.1093/nar/gku980

ImmuCo: a database of gene co-expression in immune cells

Pingzhang Wang[1,2,*,†], Huiying Qi[3,†], Shibin Song[4], Shuang Li[4], Ningyu Huang[4], Wenling Han[1,2] and Dalong Ma[1,2]

[1]Department of Immunology, Key Laboratory of Medical Immunology, Ministry of Health, School of Basic Medical Sciences, Peking University Health Science Center, No. 38 Xueyuan Road, Beijing 100191, China, [2]Peking University Center for Human Disease Genomics, No. 38 Xueyuan Road, Beijing 100191, China, [3]Department of Natural Science in Medicine, Peking University Health Science Center, No. 38 Xueyuan Road, Beijing 100191, China and [4]Information and Communication Center, Peking University Health Science Center, No. 38 Xueyuan Road, Beijing 100191, China

Received July 31, 2014; Revised September 09, 2014; Accepted October 04, 2014

ABSTRACT

Current gene co-expression databases and correlation networks do not support cell-specific analysis. Gene co-expression and expression correlation are subtly different phenomena, although both are likely to be functionally significant. Here, we report a new database, ImmuCo (http://immuco.bjmu.edu.cn), which is a cell-specific database that contains information about gene co-expression in immune cells, identifying co-expression and correlation between any two genes. The strength of co-expression of queried genes is indicated by signal values and detection calls, whereas expression correlation and strength are reflected by Pearson correlation coefficients. A scatter plot of the signal values is provided to directly illustrate the extent of co-expression and correlation. In addition, the database allows the analysis of cell-specific gene expression profile across multiple experimental conditions and can generate a list of genes that are highly correlated with the queried genes. Currently, the database covers 18 human cell groups and 10 mouse cell groups, including 20 283 human genes and 20 963 mouse genes. More than 8.6 × 10^8 and 7.4 × 10^8 probe set combinations are provided for querying each human and mouse cell group, respectively. Sample applications support the distinctive advantages of the database.

INTRODUCTION

Co-expression data are now widely used to study gene modules, gene regulation and function, protein interaction partners and signaling pathways. In addition, disease-associated gene co-expression can be used to predict tumor metastasis and patient prognosis (1–4), as well as biomarker development (5,6). Many co-expression databases have been constructed and are widely used by researchers, especially in the field of plant biology (7–13). Several co-expression databases for mammals have been established recently, including COXPRESdb (14), STARNET (15) and HGCA (16). Pearson correlation coefficients are widely used in these databases to identify gene co-expression and networks of the most highly correlated co-expressed genes. However, these databases do not support cell-specific analysis because the gene expression matrices for co-expression analysis are from multiple tissues or a mix of cells and tissues. The overall correlation in gene expression identified in these databases does not necessarily indicate that the genes co-exist in the same cell type. Actually, gene co-expression and expression correlation are subtly different phenomena, although both are likely to be functionally significant.

For wet lab experiments, more attention is paid to gene co-expression within the same tissue or cell. For example, protein interactions, cellular signaling activity and gene regulation are frequently analyzed in the same cells (such as tumor cell lines) for most experiments. Thus, correlation analysis within the same cell type no doubt provides more accurate and reliable results to guide experiments. The recently developed CHO gene co-expression database (CGCDB) (17) uses microarray data derived solely from Chinese hamster ovary (CHO) cell lines to provide cell-specific correlation analysis, but the database only contains 563 unique genes, involving 638 high confidence probe sets. Although many databases such as BioGPS (18), HemaExplorer (19), RefDIC (20), BloodExpress (21) and ImmGen (22) analyze gene expression in immune cells, they do not provide a truly direct analysis of gene co-expression or a quantitative measure of co-expression strength. In addition, the experimental conditions for the same cell types are very limited in these databases.

*To whom correspondence should be addressed. Tel: +86 10 82802846 (Ext 5036); Fax: +86 10 82801149; Email: wangpzh@bjmu.edu.cn
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Here, we report a new database, ImmuCo, which is a cell-specific database that provides co-expression analyses between any two genes in immune cells. Gene co-expression is reflected by the signal values and detection calls for a queried gene pair, whereas the strength of the expression correlation is reflected by a Pearson correlation coefficient (*r* value). ImmuCo is the first database to analyze gene co-expression independently of correlation analysis, and it is the first database to assess expression correlation in immune cells.

MATERIALS AND METHODS

Data set

Microarray data set were downloaded from the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/) (23). GEO samples related to immune cells were screened by text mining and confirmed manually (see Supplementary Methods for details).

Quality control for Affymetrix arrays

A global quality control (QC) analysis of raw data quality was performed using the BioConductor package 'simpleaffy' (24). Arrays containing extreme values from at least one QC stat were abandoned. In addition, key markers for each cell type were supposed to be expressed; that is, the detection calls for the corresponding marker probe sets should be 'present' (see Supplementary Methods for details).

Microarray analysis

Affymetrix array analysis was performed through the 'affy' package in Bio-conductor using the MAS 5.0 method (25). All default parameters, including the Chip Description File were retained. Data from each array were scaled by default to the target intensity of 500 to normalize the results for inter-array comparisons. The signal intensity value, detection *P*-value and detection call were generated for each probe set. The detection call was generated by evaluating the difference between perfect match (PM) and mismatch (MM) probe values for each probe pair in a probe set, based on the Wilcoxon's signed-rank test. Therefore, the probe sets were flagged absent (A) when the PM values were not considered to be significantly above the MM probes; otherwise, the probe sets were flagged either as present (P) or as marginally present (M) if the signal was at the limit of detection (26). For the current array platforms, those probe sets without a unique gene annotation were discarded.

Database construction

The ImmuCo database is based on Client Browser/Web Server/Database Server three-tier architecture. It is built using Apache Tomcat (web server) along with MySQL (database server). The client contains the presentation logic, including simple controls and user input validation. The web interface is built with jsp (java server pages) and follows the MVC (Model-View-Controller) development framework. The web server provides the business process logic and data access. It accepts the request and the implementation of a server-side Java programming language and returns its output, enabling the client to interact with database resources. It accesses the database using JDBC (Java Database Connectivity). The data server stores data using the MySQL RDBMS (relational database management system).

RESULTS

Data statistics

We chose the Affymetrix Human Genome U133 Plus 2.0 and Mouse Genome 430 2.0 arrays for this study because these two platforms are popular arrays with the largest available sample sizes for humans and mice, respectively, in the GEO database. After QC, 8926 human GEO samples (GSMs or microarrays) involving 344 GEO series (GSEs) and 3682 mouse samples involving 368 GSEs were retained for the gene co-expression analysis. Approximately 15% of the samples from both organisms were discarded because of quality issues. Based on sample annotation and expressed molecular markers, the selected samples were further divided into various cell types. A total of 11 human and seven mouse cell types are included in the current version of the database, including 18 human and 10 mouse cell groups (Table 1).

The human cell types include T cells, B cells, plasma cells, natural killer (NK) cells, monocytes, macrophages, dendritic cells (DCs), polymorphonuclear leukocytes (PMNs/neutrophils), peripheral blood mononuclear cells (PBMCs), hematopoietic stem cells (HSCs) and bone marrow mononuclear cells (BMMCs). The B cell, T cell and HSC groups were further divided into various groups based on the source information recorded in their SOFT format files. For example, B cells from patients with acute lymphoblastic leukaemia were placed in the 'B cell (ALL)' group, whereas B cells from patients with chronic lymphoid leukaemia were placed in the 'B cell (CLL)' group and the remaining B cells were placed into the 'B cell' group. If the T cells were a mixture of $CD4^+$ and $CD8^+$ T cells, they were grouped into the 'T cell' group; otherwise, they were placed in the CD4 ('CD4+ T cell') or CD8 ('CD8+ T cell') single positive T cell groups. The group 'T cell (ALL)' represents mixed T cells from patients with acute lymphoblastic leukaemia. Similarly, the groups 'hematopoietic stem cell (AML)', 'hematopoietic stem cell (MDS)' and 'hematopoietic stem cell' represent HSC samples from patients with acute myeloid leukaemia, myelodysplastic syndromes and patients with neither disease, respectively. Cell groups associated with disease will no doubt contribute to identifying disease-associated gene co-expression and correlations.

The mouse cell types included B cells, T cells, DCs, HSCs, macrophages, splenocytes and thymocytes. Splenocytes are actually a mixture of different white blood cell types, such as T and B lymphocytes, as long as they are situated in the spleen. Thymocytes are white blood cells situated in the thymus and primarily include T cells with distinct maturational stages based on the expression of the cell surface markers CD4 and CD8. Similar to the human 'T cell' group, the mouse 'T cell' group also contains a mixture of $CD4^+$ and $CD8^+$ T cells. The $CD4^+CD25^+$ regulatory T cells (Tregs)

Table 1. Cell types and sample size in the current version of the ImmuCo database

Species	Cell type	Sample size	GEO series number	Note
Human	AML (BMMC)	814	11	BMMCs from patients with acute myeloid leukaemia
	B cell	386	35	
	B cell (ALL)	300	7	B cells from patients with acute lymphoblastic leukaemia
	B cell (CLL)	471	12	B cells from patients with chronic lymphoid leukaemia
	$CD4^+$ T cell	551	42	
	$CD8^+$ T cell	149	22	
	DC	406	34	
	Hematopoietic stem cell (HSC)	264	39	
	Hematopoietic stem cell (AML)	113	4	HSC from patients with acute myeloid leukaemia
	Hematopoietic stem cell (MDS)	179	1	HSC from patients with myelodysplastic syndromes
	Macrophage	362	23	
	Monocyte	427	40	
	NK	128	11	
	PBMC	1921	59	
	Plasma cell	1753	12	Mainly from patients with multiple myeloma
	PMN	452	17	
	T cell	112	15	
	T cell (ALL)	138	15	T cells from patients with acute lymphoblastic leukaemia
Mouse	B cell	458	56	
	$CD4^+$ T cell	501	74	
	$CD8^+$ T cell	235	33	
	DC	347	43	
	Hematopoietic stem cell	645	86	
	Macrophage	785	58	
	Splenocyte	146	7	
	T cell	222	28	
	Thymocyte	206	20	
	Treg	137	27	

represent an important subset of helper T cells that modulate the immune system (27). Despite the large amount of public transcriptome data from immune cells, these data are largely restricted to the main categories of immune cells, such as $CD4^+$ T cells, B cells, monocytes and macrophages. Data sets describing further subsets, such as Th1, Th2 and Th17 cells, are still incomplete. Thus, in the current version of the ImmuCo database, Tregs are the only helper T cell subset represented.

Query and result description

ImmuCo provides a simple, convenient and easy-to-understand web interface that searches for and immediately calculates the results of transcriptional co-expression between any gene pair in immune cells. ImmuCo supports requests using a gene symbol or alias (e.g. RPS29 or S29), Entrez Gene ID (e.g. 6235) or probe set ID (if known, e.g. 201094_at) as the initial input (Figure 1). As the ImmuCo database focuses on the co-expression of two genes, a pair of genes must be entered in the query box to replace the default gene pair. Users can also input a single gene to retrieve the most correlated genes, and the default gene will be used as the second gene. Probe sets without a unique gene annotation were discarded, leaving 20 283 human and 20 963 mouse genes with more than 8.6×10^8 and 7.4×10^8 probe set combinations for querying each human and mouse cell group, respectively.

Features and applications

Cell-specific co-expression in samples from various experimental conditions provides a more reliable and rational explanation of gene correlation. We performed several sample applications to illustrate the database features.

(i) *Gene co-expression and correlation*. Co-expressed genes are likely to be functionally associated, and the encoded proteins may participate in the same signaling pathway, form a common structural complex, or cooperate to regulate gene expression. In the ImmuCo database, gene co-expression of queried genes is reflected by signal values and detection calls. The former indicates gene expression level, while the latter reflects co-existence state, that is, the queried gene pair shows synchronously present (present–present (PP)) or absent (absent–absent (AA)) calls in the same GEO samples. The PP (or AA) rate is calculated by dividing the sample count with PP (or AA) state by the total sample count of the queried cell group. The total co-existence rate is

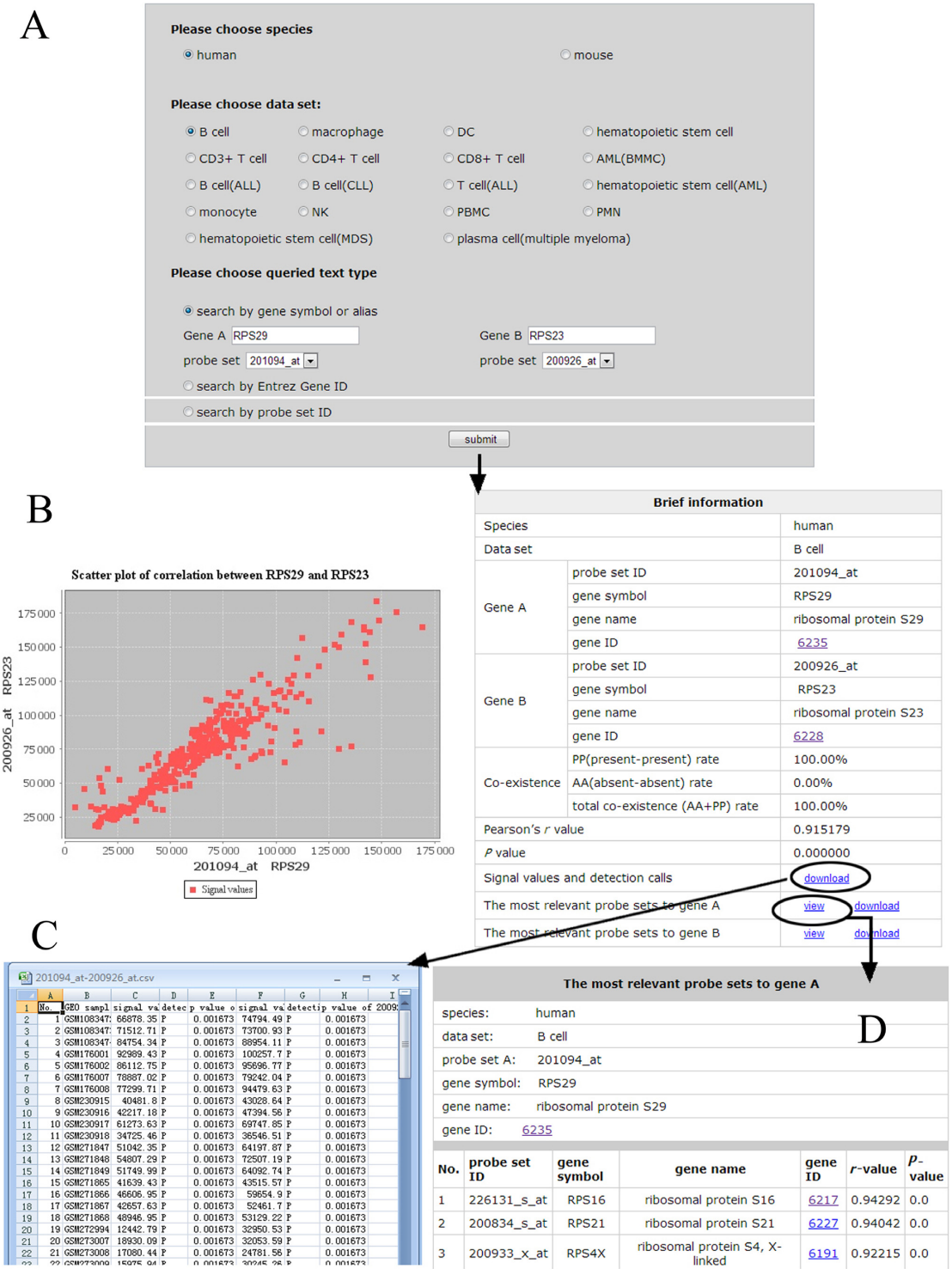

Figure 1. How to browse the ImmuCo database (the default example is shown). (**A**) Search by gene symbol or alias, which is the default option. Click the 'Gene A' or 'Gene B' textbox, and the default option disappears. A gene symbol or alias can be entered, and a corresponding probe set ID list will automatically pop up. In addition to the gene symbol, the Entrez Gene ID and the probe set ID can also be used for query types in the corresponding textboxes. (**B**) The query output. The left panel is a scatter plot of signal values for the queried gene pair. The plot directly illustrates the extent of linear correlation. In addition, co-expression of the queried genes can be identified, independently of correlation. The right panel displays information including probe set IDs, Gene IDs, HUGO gene symbols, co-existence rate, *r* value and descriptions of the queried genes and provides a download option. (**C**) GEO sample names, signal values, detection calls and *P* values can be downloaded in a CSV format file. Downloaded signal values can be used to create a similar scatter plot in Excel by user self. (**D**) The most relevant probe sets for Gene A. Currently, the 20 probe sets most correlated (based on *r* values) with Gene A or Gene B are provided for download. Gene IDs in (B) and (C) provide external links to the corresponding entries in the NCBI gene database (http://www.ncbi.nlm.nih.gov/gene). To identify co-expression relationships among multiple genes, CSV format file results can be integrated. The ImmuCo database can be accessed at its home page: http://immuco.bjmu.edu.cn.

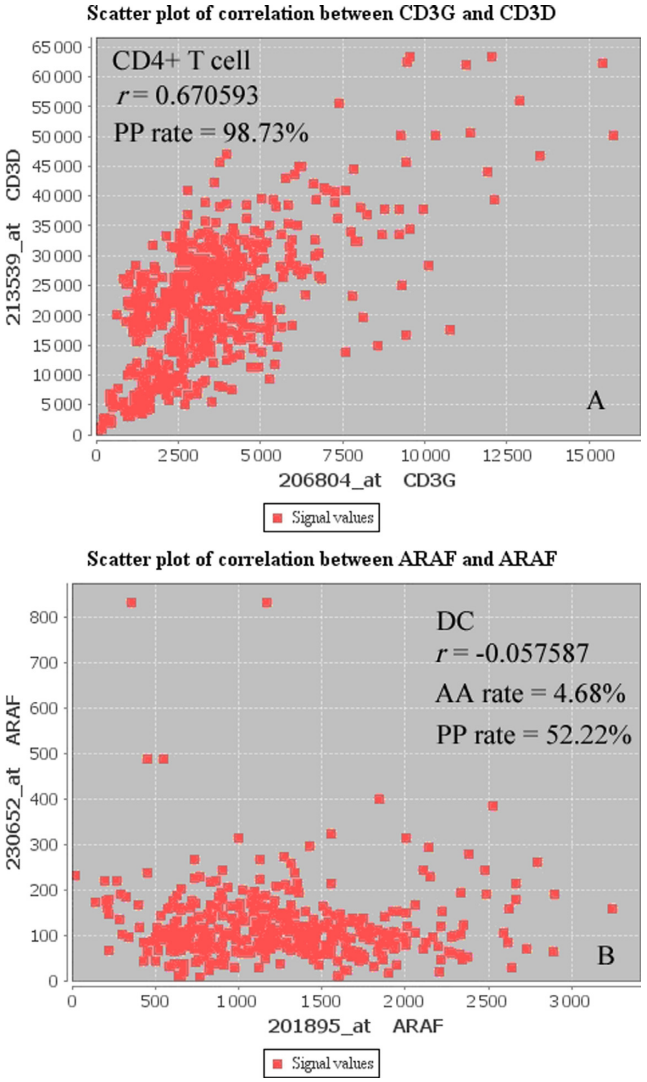

Figure 2. Sample application for gene co-expression and gene expression profile analysis. (**A**) *CD3G* and *CD3D* are significantly correlated and co-expressed in CD4$^+$ T cells. (**B**) The co-expression and correlation between probe sets for *ARAF* are shown.

the sum of the PP rate and the AA rate. For example, *CD3G* (CD3-gamma) expression is significantly correlated with *CD3D* (CD3-delta) expression in CD4$^+$ T cells (r value = 0.670593, P value = 0; PP rate = 98.7%) (Figure 2A). Both CD3G and CD3D are the components of T-cell receptor (TCR)–CD3 complex, which is expressed on the surface of T cells.

In Supplementary Figure S1 and Supplementary Table S1, we provide more examples, including membrane proteins, cytokines (secreted proteins) and transcription factors (nuclear proteins) in CD4$^+$ T cells, to illustrate the co-expression and correlation analysis and to support the cell-specific advantage. For example, *CD3E* (CD3-epsilon) and *CD247* (both encode components of TCR–CD3 complex) are co-expressed (PP rate = 98.7%) but not well-correlated at the mRNA level (r value = 0.087085, P value = 0.041001). These results suggest that the profiles of functionally associated genes, even if they encode components of the same protein complex, are not necessarily correlated. The terms 'correlated' and 'co-expressed' therefore cannot be considered conceptually equal. Therefore, ImmuCo provides a quick view of gene co-expression, correlation (both positive and negative correlations) and the strength of co-expression and correlation, and can detect the subtle difference between co-expression and correlation.

(ii) *Cell-specific gene expression profile analysis*. The ImmuCo database provides a global, two-dimensional view of cell-specific signal values, which constitute a single gene expression profile on either axis of the scatter plot, across different experimental conditions. The scatter plot provides important information regarding the general expression level of a gene. For example, IL-4, IL-5 and IL-13 are generally expressed at low levels in most CD4$^+$ T cells, and the data points cluster at the origin of the graph (Supplementary Figure S1).

Gene expression profiles of different transcript variants of the same gene can also be illustrated, but these transcript variants are mainly from different polyadenylation sites because the probes are designed mainly in the 3′-untranslated region. For example, human *ARAF* (serine/threonine-protein kinase A-Raf) has two probe sets (201895_at and 230652_at), but they correspond to different transcript variants. The former corresponds to ARAF transcript variant 1 (NM_001654) and 2 (NM_001256196), while the latter corresponds to ARAF transcript variant 3 (NM_001256197), which results from an alternative intronic polyadenylation site (28). As shown in Figure 2B, these two probe sets are not highly correlated (r value = −0.057587, P value = 0.246972), but they can co-exist in over half of samples in DCs (the total co-existence (PP+AA) rate = 56.9%), though the expression level of transcript variant 3 is actually very low under a series of experimental conditions.

Low signal values in a profile often indicate no or low expression levels, but sometimes poor probe quality should also be considered (Supplementary Figure S2). The compromise reference thresholds of signal value for gene expression can be set at 150 and 100 for human and mouse, respectively (see the Discussion section).

DISCUSSION

Our newly established ImmuCo database provides a simple and effective way to identify co-expression and correlation between any gene pair under a series of experimental conditions. This goal cannot be achieved using other existing mammalian gene expression or co-expression databases (14–16,18–22). Traditionally, standard methods such as the Pearson and Spearman correlations are used to identify gene co-expression and correlation relationships. However, co-expression and expression correlation are subtly different phenomena under many conditions. For microarray data, co-expression does not necessarily indicate expression correlation and vice versa. One reason for this distinction is high levels of background noise, which may result in genes appearing to be correlated when they are not

co-expressed. In contrast, parallel co-expression networks are involved in certain cellular processes, but the individual genes that make up these networks may not be correlated at the mRNA level. In addition, negative correlated genes can be either co-expressed or mutually exclusive. Therefore, databases such as COXPRESdb (14,29), STARNET (15,30) and HGCA (16) mainly provide information about positive gene correlation and do not reflect co-expression relationships between genes with no apparent correlation.

In the ImmuCo database, co-expression is indicated by signal values and detection calls. The signal values reflect the relative expression levels of queried gene pair. Both positive and negative correlations can be reflected by the tendency of the change of signal values. The signal values derived from the MAS5 algorithm are usually normalized using the signal intensities of the PM probes subtracted by the MM probes, but the MM values may not ideally represent the background. The target intensity of all arrays is set at 500; very low signal values on either axis of the signal scatter plot indicate to the user that the gene may be not expressed, whereas large signal values often indicate that the gene is expressed. The values 100 and 270 represent approximately the third quartile (75th percentile, 3rd Qu./Q3) and the first quartile (25th percentile, 1st Qu./Q1) of the values of all probe sets with absent and present calls, respectively, in all human cells, while the corresponding values for mouse cells are about 55 (Q3) and 200 (Q1) for the probe sets with absent and present calls, respectively. It is difficult to set a unified threshold to judge the expression state for all probe sets. The compromise values, 150 for human and 100 for mouse, could exclude about 90% probe sets with absent calls, though about 10% probe sets with present calls were also excluded. Therefore, these two values may be used as reference thresholds to judge gene expression for human and mouse, respectively.

It is important to note that the most highly correlated gene sets for a single queried gene vary considerably between the current co-expression databases (Supplementary Table S1). The inconsistency can be attributed to high levels of background noise in microarray data, variable cellular states, cell- and tissue-specific expression, different methods of correlation analysis and other factors, even though the same microarray platforms, such as Affymetrix Human Genome U133 plus 2.0 arrays, are commonly used in these databases. Further investigation is needed to find the key determinants that result in this inconsistency and to identify the characteristics of common genes with conserved co-expression or correlation patterns. Thus, a systematic comparison should be performed to address these questions.

In addition to gene co-expression and correlation, the ImmuCo database provides a direct global view of gene expression values across various conditions in each cell type. Because the immune cells used to generate the gene expression data are from different individuals, different physiological and pathological states and various experimental or treatment conditions, the scatter plot shows that the expression levels of a single gene are highly variable, suggesting high dynamic and plasticity in gene expression. Our established database, ImmuSort (http://202.85.212.211/Account/ImmuSort.html; submitted), is designed to highlight gene plasticity, which is defined as the extent of change in gene expression in response to various environmental or genetic influences. In addition, the ImmuSort database electronically sorts gene expression intensity data by the experimental conditions and cell states associated with a certain expression level. ImmuSort provides the comparison of gene expression intensity of different transcripts at probe set levels, therefore, it can help users to choose suitable probe sets for co-expression analysis.

We are planning to integrate cell-specific gene co-expression network graphics into future versions of the ImmuCo database, similar to other co-expression-related databases. In addition, GO annotations (31), KEGG pathways (32) and protein–protein interaction data from other databases, such as HPRD (33) and IntACT (34), may also be incorporated to further enrich the database content. Moreover, non-immune cells will also be integrated into an expanded version of ImmuCo. Because gene correlation and co-expression at the mRNA level do not necessarily indicate expression correlations at the protein level, and because not all genes are strictly regulated at the mRNA level, there is still a need for databases that analyze co-expression at the protein level.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Natural Science Foundation of China [31270948]; Program for Innovation of New Drugs [2013ZX09103003-023]; Beijing Natural Science Foundation [5122016]. Funding for open access charge: Beijing Natural Science Foundation [5122016].
Conflict of interest statement. None declared.

REFERENCES

1. Futamura,N., Nishida,Y., Urakawa,H., Kozawa,E., Ikuta,K., Hamada,S. and Ishiguro,N. (2014) EMMPRIN co-expressed with matrix metalloproteinases predicts poor prognosis in patients with osteosarcoma. *Tumour Biol.*, **35**, 5159–5165.
2. Ma,R.L., Shen,L.Y. and Chen,K.N. (2014) Coexpression of ANXA2, SOD2 and HOXA13 predicts poor prognosis of esophageal squamous cell carcinoma. *Oncol. Rep.*, **31**, 2157–2164.
3. Puzovic,V., Brcic,I., Ranogajec,I. and Jakic-Razumovic,J. (2014) Prognostic values of ETS-1, MMP-2 and MMP-9 expression and co-expression in breast cancer patients. *Neoplasma*, **61**, 439–447.
4. Huang,S., Zhong,X., Gao,J., Song,R., Wu,H., Zi,S., Yang,S., Du,P., Cui,L., Yang,C. *et al.* (2014) Coexpression of SFRP1 and WIF1 as a prognostic predictor of favorable outcomes in patients with colorectal carcinoma. *Biomed. Res. Int.*, **2014**, doi:10.1155/2014/256723.
5. Erbel,C., Tyka,M., Helmes,C.M., Akhavanpoor,M., Rupp,G., Domschke,G., Linden,F., Wolf,A., Doesch,A., Lasitschka,F. *et al.* (2014) CXCL4-induced plaque macrophages can be specifically identified by co-expression of MMP7+S100A8+ in vitro and in vivo. *Innate Immun.*, doi:10.1177/1753425914526461.
6. Sun,Y., Zhang,W., Chen,D., Lv,Y., Zheng,J., Lilljebjorn,H., Ran,L., Bao,Z., Soneson,C., Sjogren,H.O. *et al.* (2014) A glioma classification scheme based on coexpression modules of EGFR and PDGFRA. *Proc. Natl Acad. Sci. U.S.A.*, **111**, 3538–3543.
7. Manfield,I.W., Jen,C.H., Pinney,J.W., Michalopoulos,I., Bradford,J.R., Gilmartin,P.M. and Westhead,D.R. (2006) Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis. *Nucleic Acids Res.*, **34**, W504–W509.

8. Obayashi,T., Okamura,Y., Ito,S., Tadaka,S., Aoki,Y., Shirota,M. and Kinoshita,K. (2014) ATTED-II in 2014: evaluation of gene coexpression in agriculturally important plants. *Plant Cell Physiol.*, **55**, e6.
9. Wong,D.C., Sweetman,C., Drew,D.P. and Ford,C.M. (2013) VTCdb: a gene co-expression database for the crop species Vitis vinifera (grapevine). *BMC Genomics*, **14**, 882.
10. Sato,Y., Namiki,N., Takehisa,H., Kamatsuki,K., Minami,H., Ikawa,H., Ohyanagi,H., Sugimoto,K., Itoh,J., Antonio,B.A. *et al.* (2013) RiceFREND: a platform for retrieving coexpressed gene networks in rice. *Nucleic Acids Res.*, **41**, D1214–D1221.
11. Yim,W.C., Yu,Y., Song,K., Jang,C.S. and Lee,B.M. (2013) PLANEX: the plant co-expression database. *BMC Plant Biol.*, **13**, 83.
12. Ogata,Y., Suzuki,H., Sakurai,N. and Shibata,D. (2010) CoP: a database for characterizing co-expressed gene modules with biological information in plants. *Bioinformatics*, **26**, 1267–1268.
13. De Bodt,S., Hollunder,J., Nelissen,H., Meulemeester,N. and Inze,D. (2012) CORNET 2.0: integrating plant coexpression, protein-protein interactions, regulatory interactions, gene associations and functional annotations. *New Phytol.*, **195**, 707–720.
14. Obayashi,T., Okamura,Y., Ito,S., Tadaka,S., Motoike,I.N. and Kinoshita,K. (2013) COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals. *Nucleic Acids Res.*, **41**, D1014–D1020.
15. Jupiter,D., Chen,H. and VanBuren,V. (2009) STARNET 2: a web-based tool for accelerating discovery of gene regulatory networks using microarray co-expression data. *BMC Bioinformatics*, **10**, 332.
16. Michalopoulos,I., Pavlopoulos,G.A., Malatras,A., Karelas,A., Kostadima,M.A., Schneider,R. and Kossida,S. (2012) Human gene correlation analysis (HGCA): a tool for the identification of transcriptionally co-expressed genes. *BMC Res. Notes*, **5**, 265.
17. Clarke,C., Doolan,P., Barron,N., Meleady,P., Madden,S.F., DiNino,D., Leonard,M. and Clynes,M. (2012) CGCDB: a web-based resource for the investigation of gene coexpression in CHO cell culture. *Biotechnol. Bioeng.*, **109**, 1368–1370.
18. Wu,C., Orozco,C., Boyer,J., Leglise,M., Goodale,J., Batalov,S., Hodge,C.L., Haase,J., Janes,J., Huss,J.R. *et al.* (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.*, **10**, R130.
19. Bagger,F.O., Rapin,N., Theilgaard-Monch,K., Kaczkowski,B., Thoren,L.A., Jendholm,J., Winther,O. and Porse,B.T. (2013) HemaExplorer: a database of mRNA expression profiles in normal and malignant haematopoiesis. *Nucleic Acids Res.*, **41**, D1034–D1039.
20. Hijikata,A., Kitamura,H., Kimura,Y., Yokoyama,R., Aiba,Y., Bao,Y., Fujita,S., Hase,K., Hori,S., Ishii,Y. *et al.* (2007) Construction of an open-access database that integrates cross-reference information from the transcriptome and proteome of immune cells. *Bioinformatics*, **23**, 2934–2941.
21. Miranda-Saavedra,D., De,S., Trotter,M.W., Teichmann,S.A. and Gottgens,B. (2009) BloodExpress: a database of gene expression in mouse haematopoiesis. *Nucleic Acids Res.*, **37**, D873–D879.
22. Shay,T. and Kang,J. (2013) Immunological Genome Project and systems immunology. *Trends Immunol.*, **34**, 602–609.
23. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
24. Wilson,C.L. and Miller,C.J. (2005) Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. *Bioinformatics*, **21**, 3683–3685.
25. Gautier,L., Cope,L., Bolstad,B.M. and Irizarry,R.A. (2004) affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.
26. Liu,W.M., Mei,R., Di,X., Ryder,T.B., Hubbell,E., Dee,S., Webster,T.A., Harrington,C.A., Ho,M.H., Baid,J. *et al.* (2002) Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics*, **18**, 1593–1599.
27. Sakaguchi,S., Yamaguchi,T., Nomura,T. and Ono,M. (2008) Regulatory T cells and immune tolerance. *Cell*, **133**, 775–787.
28. Wang,P., Yu,P., Gao,P., Shi,T. and Ma,D. (2009) Discovery of novel human transcript variants by analysis of intronic single-block EST with polyadenylation site. *BMC Genomics*, **10**, 518.
29. Obayashi,T., Hayashi,S., Shibaoka,M., Saeki,M., Ohta,H. and Kinoshita,K. (2008) COXPRESdb: a database of coexpressed gene networks in mammals. *Nucleic Acids Res.*, **36**, D77–D82.
30. Jupiter,D.C. and VanBuren,V. (2008) A visual data mining tool that facilitates reconstruction of transcription regulatory networks. *PLoS One*, **3**, e1717.
31. Blake,J.A., Dolan,M., Drabkin,H., Hill,D.P., Li,N., Sitnikov,D., Bridges,S., Burgess,S., Buza,T., McCarthy,F. *et al.* (2013) Gene Ontology annotations and resources. *Nucleic Acids Res.*, **41**, D530–D535.
32. Kanehisa,M., Goto,S., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.
33. Keshava,P.T., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database-2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
34. Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., Del-Toro,N. *et al.* (2014) The MIntAct project-IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.