OXFORD

Briefings in Bioinformatics, 2017, 1–10

doi: 10.1093/bib/bbx024
Paper

Comparison of the general co-expression landscapes between human and mouse

Di Liu,* Linna Zhao,* Yang Chen,* Zhaoyang Wang,* Jing Xu,* Ying Li, Changgui Lei, Simeng Hu, Miaomiao Niu and Yongshuai Jiang

Corresponding author: Yongshuai Jiang, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China, and Training Center for Students Innovation and Entrepreneurship Education, Harbin Medical University, Harbin 150086, China. E-mail: jiangyongshuai@gmail.com or jiangyongshuai@ems.hrbmu.edu.cn
*These authors contributed equally to this work.

Abstract

The murine model serves as an important experimental system in biomedical science because of its high degree of similarities at the sequence level with human. Recent studies have compared the transcriptional landscapes between human and mouse, but the general co-expression landscapes have not been characterized. Here, we calculated the general co-expression coefficients and constructed the general co-expression maps for human and mouse. The differences and similarities of the general co-expression maps between the two species were compared in detail. The results showed low similarities in the human and mouse, with only about 36.54% of the co-expression relationships conserved between the two species. These results indicate that researchers should pay attention to these differences when performing research using the expression data of human and mouse. To facilitate use of this information, we also developed the human–mouse general co-expression difference database (coexpressMAP) to search differences in co-expression between human and mouse. This database is freely available at http://www.bioapp.org/coexpressMAP.

Key words: general co-expression; orthologous gene; human; mouse

Di Liu is a master at College of Bioinformatics Science and Technology, Harbin Medical University. She is also a member of genetic testing group at Training Center for Students Innovation and Entrepreneurship Education of Harbin Medical University. Her research interest focuses on statistical genetics and bioinformatics.
Linna Zhao is a master at College of Bioinformatics Science and Technology, Harbin Medical University. She is also a member of genetic testing group at Training Center for Students Innovation and Entrepreneurship Education of Harbin Medical University. Her research interest focuses on statistical genetics and bioinformatics.
Yang Chen is a bachelor at College of Bioinformatics Science and Technology, Harbin Medical University. His research interest focuses on statistical genetics.
Zhaoyang Wang is a bachelor at College of Bioinformatics Science and Technology, Harbin Medical University. Her research interest focuses on statistical genetics.
Jing Xu is a bachelor at College of Bioinformatics Science and Technology, Harbin Medical University. Her research interest focuses on statistical genetics.
Ying Li is a bachelor training at College of Bioinformatics Science and Technology, Harbin Medical University. She is also a member of genetic testing group at Training Center for Students Innovation and Entrepreneurship Education of Harbin Medical University. Her research interest focuses on statistical genetics.
Changgui Lei is a bachelor at College of Bioinformatics Science and Technology, Harbin Medical University. His research interest focuses on statistical genetics.
Simeng Hu is a bachelor at College of Bioinformatics Science and Technology, Harbin Medical University. Her research interest focuses on statistical genetics.
Miaomiao Niu is a bachelor at College of Bioinformatics Science and Technology, Harbin Medical University. Her research interest focuses on statistical genetics.
Yongshuai Jiang is an associate professor at College of Bioinformatics Science and Technology, Harbin Medical University. He is the leader of genetic testing group at Training Center for Students Innovation and Entrepreneurship Education of Harbin Medical University. His research interest focuses on bioinformatics and population genetics.
Submitted: 5 December 2016; **Received (in revised form):** 10 February 2017

© The Author 2017. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Introduction

The mouse is an important laboratory model species that has been widely used in the research of human diseases, testing drug responses and biology [1–6]. The major hypothesis underlying its widespread use in research is that fundamental biochemical and cellular processes are conserved between human and mouse. Although a high degree of similarities has been demonstrated at the sequence level (70.1% of the residues are identical for human–mouse 1:1 orthologous, and the median amino acid identity is 78.5%) [7], there are still many differences between the two species.

A recent study compared the transcriptional landscapes between human and mouse tissues by examining the expression profiles of 15 tissues and found more differences than similarities between the two species [8]. Tsaparas *et al.* [9] constructed and compared the co-expression networks between human and mouse using 28 tissue samples. In 2015, Monaco *et al.* [10] determined the frequency in which expressions of two genes were simultaneously increased or decreased across different conditions for different data sets and constructed human and mouse co-expression networks using the top 5% co-expressed genes. Although these studies have achieved great success and identified some differences between human and mouse, they considered only a small portion (about 5%) of the co-expression relationships, and thus have provided limited understanding of the human–mouse co-expression differences. Until now, the global differences and similarities of the general co-expression landscapes still have not been characterized.

In this study, the gene co-expression relationships were classified into two categories: special co-expression relationships and general co-expression relationships. A special co-expression relationship was defined as the simultaneous expression of two or more genes in cells or tissues under a specific cell cycle state, developmental stage or external signal [11–13]. For example, Liu *et al.* [14] analyzed special co-expression relationships at different human cell development stages and found some developmental stage-specific co-expression modules; for example, they found two co-expression modules for the oocyte stage, one module for the zygote stage, one module for the two-cell stage and five modules for the four-cell stage. The special co-expression relationships will fluctuate with different sample types or under different given biological conditions. When these specific conditions are relaxed and the cell or tissue type, cycle state or developmental stage are not considered, the co-expression relationships will reflect the collaboration between two genes in a species. We define this relationship as a general co-expression relationship. If two genes have a stronger general co-expression relationship, they will be co-expressed in most biological conditions. In other words, the two genes have stronger association. The special co-expression relationships will fluctuate with different biological conditions, while the general co-expression relationship will be stable across the samples in a species. The general co-expression relationship is an inherent property of a gene pair in a species, and thus can be compared among different species.

In this study, we first described the stability of the general co-expression relationships and then compared the general co-expression landscapes between human and mouse using normal (healthy) samples. We found a large difference in general co-expression relationships between human and mouse. To facilitate inquiry into the differences and similarities of general co-expression relationships between the two species, we developed a free online database coexpressMAP: the human–mouse general co-expression differences database. The coexpressMAP can be accessed at http://www.bioapp.org/coexpressMAP.

Results

To compare the differences and similarities of the general co-expression landscapes between human and mouse in greater detail, we built a matrix of a human–mouse one-to-one orthologous gene expression profile that consists of 14 331 rows (one-to-one orthologous gene pairs) and 6032 columns (3671 human samples and 2361 mouse samples).

The stability and repeatability of the general co-expression relationships

For a given pair of genes, we examined whether the general co-expression relationship is stable. As an example, we randomly selected five pairs of genes from 14 331 genes (seed = 5), and calculated the general co-expression coefficients (GCCs, see Methods section) using different sample sizes (the samples were also randomly selected from each of the human or mouse sample set). Both Figure 1A (human) and B (mouse) displayed that, for each gene pair, GCCs tend to be stable when the sample size is increased. Furthermore, we examined how many samples are needed when we repeat the experiment. We randomly selected five groups of genes (100 genes in each group). For each gene pair in each group, we calculated the GCC values using randomly selected samples (the sample size ranged from 1 to 1000). To examine the repeatability, we selected the same number of samples from the remaining samples and recalculated the GCC values. For each of the five groups, the Pearson correlation coefficients between the first experiment and second replication are shown in Figure 1C (human) and D (mouse). In both human and mouse, the Pearson correlation coefficients increased with the increase of sample size. In fact, most of the Pearson correlation coefficients were >0.98 when the sample size was >500. In other words, the GCC is stable and repeatable when the sample size is relatively large. In subsequent analysis, we calculated the GCC for two complete data sets H (3671 human samples) and M (2361 mouse samples) to compare the differences and similarities between human and mouse. We also constructed four independent data sets: H1 (1000 human samples), H2 (1000 human samples), M1 (1000 mouse samples) and M2 (1000 mouse samples). H2 and M2 were used to repeat the results derived from H1 and M1.

Comparison of the distribution of GCCs between human and mouse

For each of two complete sample sets (H and M) and four independent sample sets (H1, H2, M1 and M2), we calculated 102 681 615 (C^2_{14331}) GCC values for all the pair-wise genes. The distribution of GCCs is shown in Figure 2A. We observed similar distribution shapes between human and mouse. In both human and mouse, the distributions of GCC values were not normal and showed a left-skewed bias. Most of the co-expression relationships between genes showed a positive correlation (the GCC of a pair-wise gene >0).

We did observe some differences between human and mouse. The skewness of human GCCs (-0.813 for H1 and -0.848 for H2) was larger than those of mouse (-0.589 for M1 and -0.575 for M2; Figure 2B). A similar phenomenon can also be observed for kurtosis, mean and median (Figure 2B; for detailed

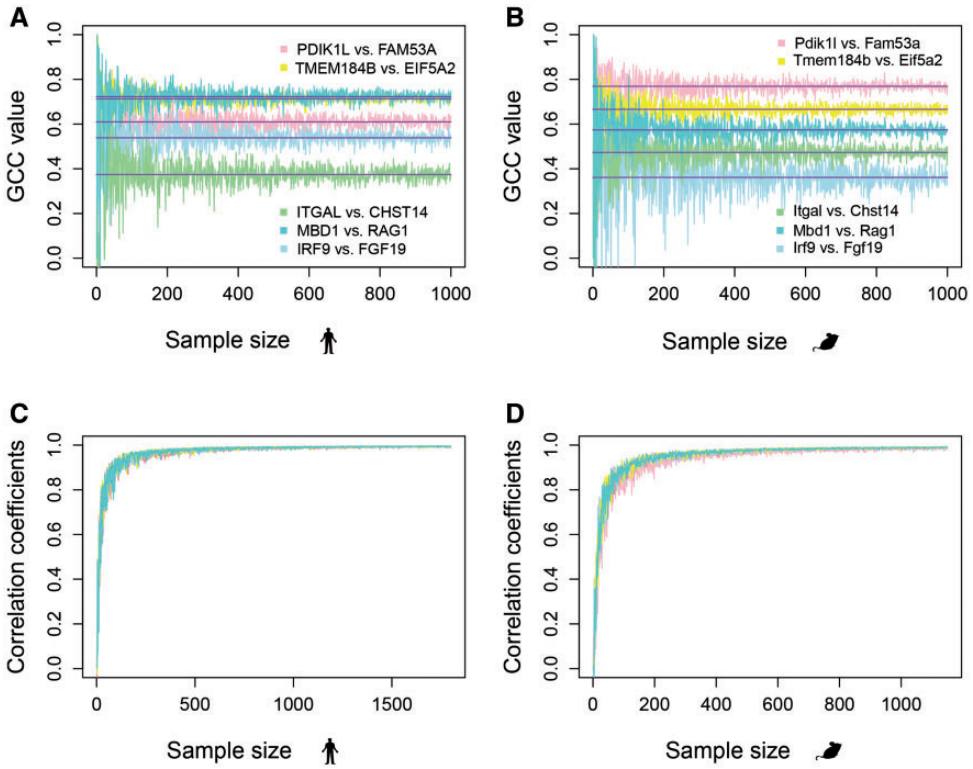

Figure 1. The stability and the repeatability of GCC. (**A**) The relationship between the sample size and the stability of GCC values for the five gene pairs in human; (**B**) the relationship between the sample size and the stability of GCC values for the five gene pairs in mouse; (**C**) the relationship between sample size and repeatability of the general co-expression relationships for the five groups (100 genes in each group) of genes in human; (**D**) the relationship between sample size and repeatability of the general co-expression relationships for the five groups (100 genes in each group) of genes in mouse.

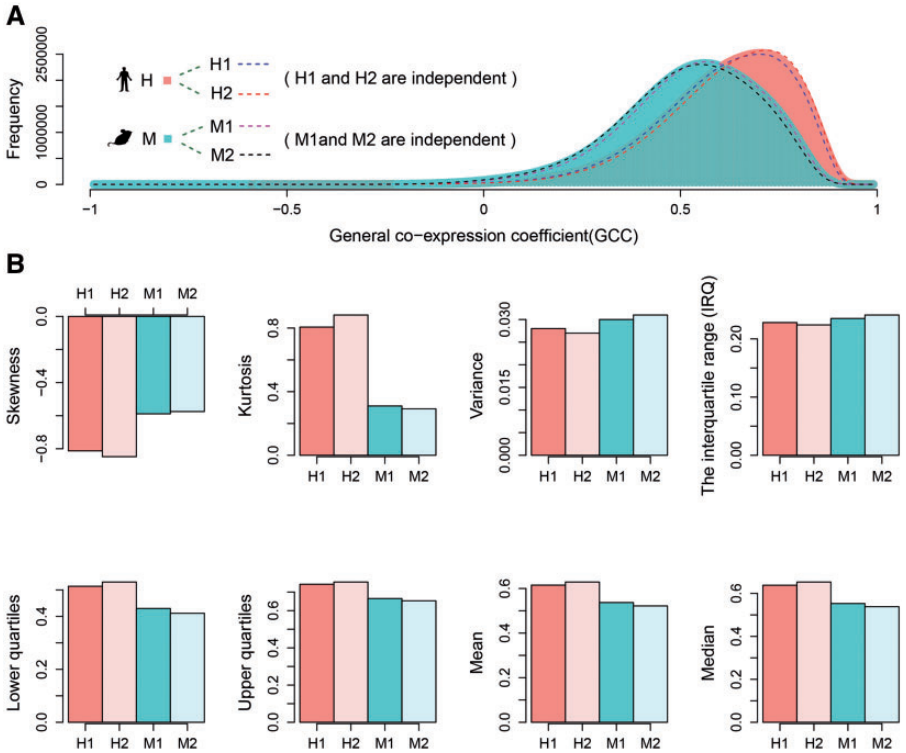

Figure 2. The distribution of GCC. (**A**) The distribution of GCC values in human (H, H1 and H2) and mouse (M, M1 and M2); (**B**) the skewness, kurtosis, variance, the interquartile range (IQR), lower quartiles, upper quartiles, mean and median for H1, H2, M1 and M2.

information, see Table 1). These data indicated that the degree of gene co-expression was higher in human than mouse.

Only 36.54% of gene pairs maintain a robust co-expression relationship

We next analyzed the changes of GCCs between human and mouse. We calculated the GCC differences between H1, H2, M1 and M2 (H1-H2, H1-M1, H1-M2, H2-M1, H2-M2 and M1-M2). The distributions of differences between any two sets are shown in Figure 3A–F. We observed fewer differences within the same species than between species. Figure 3C (mouse) and 3D (human) shows the distributions of GCC differences within the same species. These differences are mainly random errors, which are caused by random sampling. The means of these differences are approximate to 0 (Figure 3C and D), and SDs are approximate to 0.0265 (0.024 for H1-H2 and 0.029 for M1-M2). Here, we used triple SD (−0.0795, 0.0795) to represent the range of error (99% reference value range). This range was used to analyze the conserved (or robust) co-expression relationships between human and mouse. That is, if the difference value of a GCC between human and mouse falls into the range (−0.0795, 0.0795), we believe that the co-expression relationship is not changed. In other words, for robust co-expression relationships, the different GCC values between human and mouse are caused by random sampling. For the two complete data sets H and M, we also draw the distributions of GCC differences (Supplementary Figure S1). We then scanned all 102 681 615 gene pairs and calculated the percentage of conserved co-expression relationships. We found that only about 36.54% of the co-expression relationships calculated using H and M were conserved between human and mouse (Table 2).

We also counted the number of conserved co-expression relationships at each GCC level. From Figure 4, we can see that the conserved co-expression relationships also showed a unimodal distribution. For the same species (Figure 4A and F), the curves of conserved co-expression relationships were consistent with the curves of general co-expression relationships. The percentages of conserved co-expression relationships were >90% for most of the GCC levels (Figure 4G and L). This indicated that the general co-expression relationships were conserved within the same species. For different species (Figure 4B–E), the curves of conserved co-expression relationships were much lower than those of general co-expression relationships. We observed that the percentages of conserved co-expression relationships were not uniform (Figure 4H–K). The percentage curves of conserved co-expression relationships were first increased, then decreased and finally reached the maximum at GCC > 0.98. This indicated that, for the most part, the general co-expression relationships were not conserved, and only in rare situations (such as GCC in 0.98–1), the general co-expression relationships displayed high evolutionary conservation.

Table 1. Compare the distribution between human and mouse

Character of distribution	H	M	H1	H2	M1	M2
Skewness	−0.834	−0.587	−0.813	−0.848	−0.589	−0.575
Kurtosis	0.878	0.306	0.806	0.882	0.310	0.292
Mean	0.618	0.533	0.615	0.629	0.537	0.522
Variance	0.028	0.031	0.028	0.027	0.030	0.031
Minimum	−0.588	−0.485	−0.583	−0.580	−0.481	−0.496
Lower quartiles	0.518	0.424	0.514	0.530	0.430	0.412
Median	0.641	0.549	0.638	0.653	0.553	0.538
Upper quartiles	0.744	0.663	0.742	0.754	0.665	0.653
Maximum	1.000	1.000	1.000	1.000	1.000	1.000
IQR	0.226	0.239	0.228	0.224	0.235	0.241

Analysis of largely changed general co-expression relationships

As shown in Figure 3A, B, E and F, and Supplementary Figure S1, most of the changes of general co-expression relationships between human and mouse are beyond the allowed range of random errors. It is noteworthy that some larger changes of GCC between human and mouse are >0.5. The gene pair SLC27A5/Slc27a5-SLC25A4/Slc25a4 has the maximum change of GCC. In human, the GCC of *SLC27A5-SLC25A4* is 0.779 for H (0.79 for H1 and 0.779 for H2), whereas in the mouse, the GCC of *Slc27a5-Slc25a4* is −0.374 for M (−0.364 for M1 and −0.381 for M2). The differences in GCC between human and mouse were >1.0 (1.153 for H and M, 1.154 for H1-M1, 1.171 for H1-M2, 1.143 for H2-M1 and 1.160 for H2-M2). For the largely changed general co-expression relationships between human and mouse, we also counted their number and percentage at each GCC level ([−1,1] was divided into 100 levels). These gene pairs showed lower co-expression relationships in mouse; however, they showed a higher co-expression relationship in human (Figure 5A–E). In the mouse genome, largely changed general co-expression relationships had a higher percentage at negative GCC levels (Figure 5F–J). This indicated that, for a gene pair, the higher the degree of negative correlation, the more likely it is to have a large change in the co-expression relationship. We also observed that some gene pairs showed lower co-expression relationships in human; however, they showed a higher co-expression relationship in mouse. But the number of such gene pairs is relatively small. More detailed results can be seen in Supplementary Figure S2.

Comparison of the general co-expression network between human and mouse

For co-expression relationships with higher GCC values, we also used the network to compare the differences and similarities between human and mouse. We first sorted the GCC values of all 102 681 615 gene pairs for both human and mouse. We then constructed the human general co-expression network (HGCN) and mouse general co-expression network (MGCN) using the top 1 000 000 co-expression relationships. To illustrate the reliability of the results, we also used the four sample sets H1, H2, M1 and M2 to build four networks: HGCN1 (10 573 nodes), HGCN2 (10 584 nodes), MGCN1 (10 669 nodes) and MGCN2 (10 585 nodes). Four network topology characteristics (degree, betweenness, clustering coefficient and shortest path length) were used to compare the differences and similarities between human and mouse. Figure 6 shows that all the degrees of the four networks obey approximately power-law distribution (r^2=0.83 for HGCN1, r^2=0.843 for HGCN2, r^2=0.852 for MGCN1 and r^2=0.841 for MGCN2). However, the correlations of degree between human and mouse were lower (the Pearson correlation coefficient is 0.279 for H1 and M1; 0.265 for H1 and M2; 0.267 for H2 and M1; and 0.258 for H2 and M2). The same phenomenon can also be observed for betweenness, clustering coefficient and shortest path length. The human and mouse have similar distribution shapes for betweenness, clustering coefficient and shortest path length; however, the genes themselves showed lower correlation (Supplementary Figure S3 for betweenness; Supplementary Figure S4 for clustering coefficient; and Supplementary Figure S5 for shortest path length). These

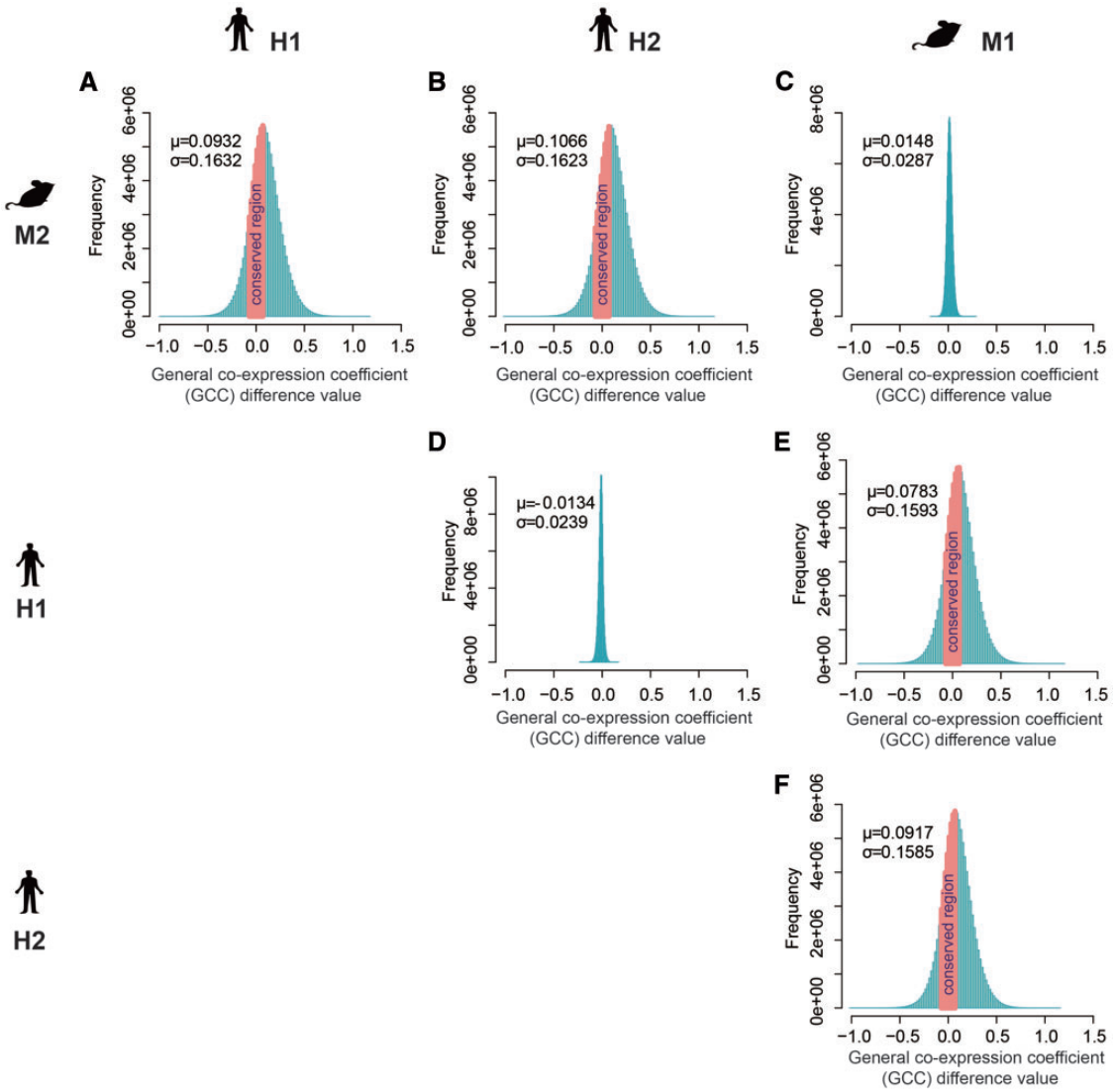

Figure 3. The distribution of GCC differences between human and mouse. (**A**) The distribution of GCC differences between H1 and M2; (**B**) the distribution of GCC differences between H2 and M2; (**C**) the distribution of GCC differences between M1 and M2; (**D**) the distribution of GCC differences between H1 and H2; (**E**) the distribution of GCC differences between H1 and M1; (**F**) the distribution of GCC differences between H2 and M1.

Table 2. The percentage of the robust co-expression relationships

Type	Set pair	Total gene pairs	Robust gene pairs	Percentage
Within species	H1-H2	102 681 615	101 705 904	99.05
Within species	M1-M2	102 681 615	100 738 682	98.11
Average				98.58
Between species	H1-M1	102 681 615	38 318 106	37.32
Between species	H1-M2	102 681 615	35 787 977	34.85
Between species	H2-M1	102 681 615	36 545 838	35.59
Between species	H2-M2	102 681 615	33 909 021	33.02
Between species	H-M	102 681 615	37 518 602	36.54

results indicated that although the human and mouse had similar global network properties, they displayed great differences in local network properties. This conclusion is similar to the co-expression network constructed using 28 samples by Tsaparas *et al.* [9]. We also compared the topology characteristics of general co-expression network for the same species (HGCN1 versus HGCN2 and MGCN1 versus MGCN2). We observed a high degree of consistency for both global and local network properties (Figure 6 for degree; Supplementary Figure S3 for betweenness; Supplementary Figure S4 for clustering coefficient; and Supplementary Figure S5 for shortest path length). This also indicated that the general co-expression relationships were stable within a species, and the differences dominate similarities between the two species.

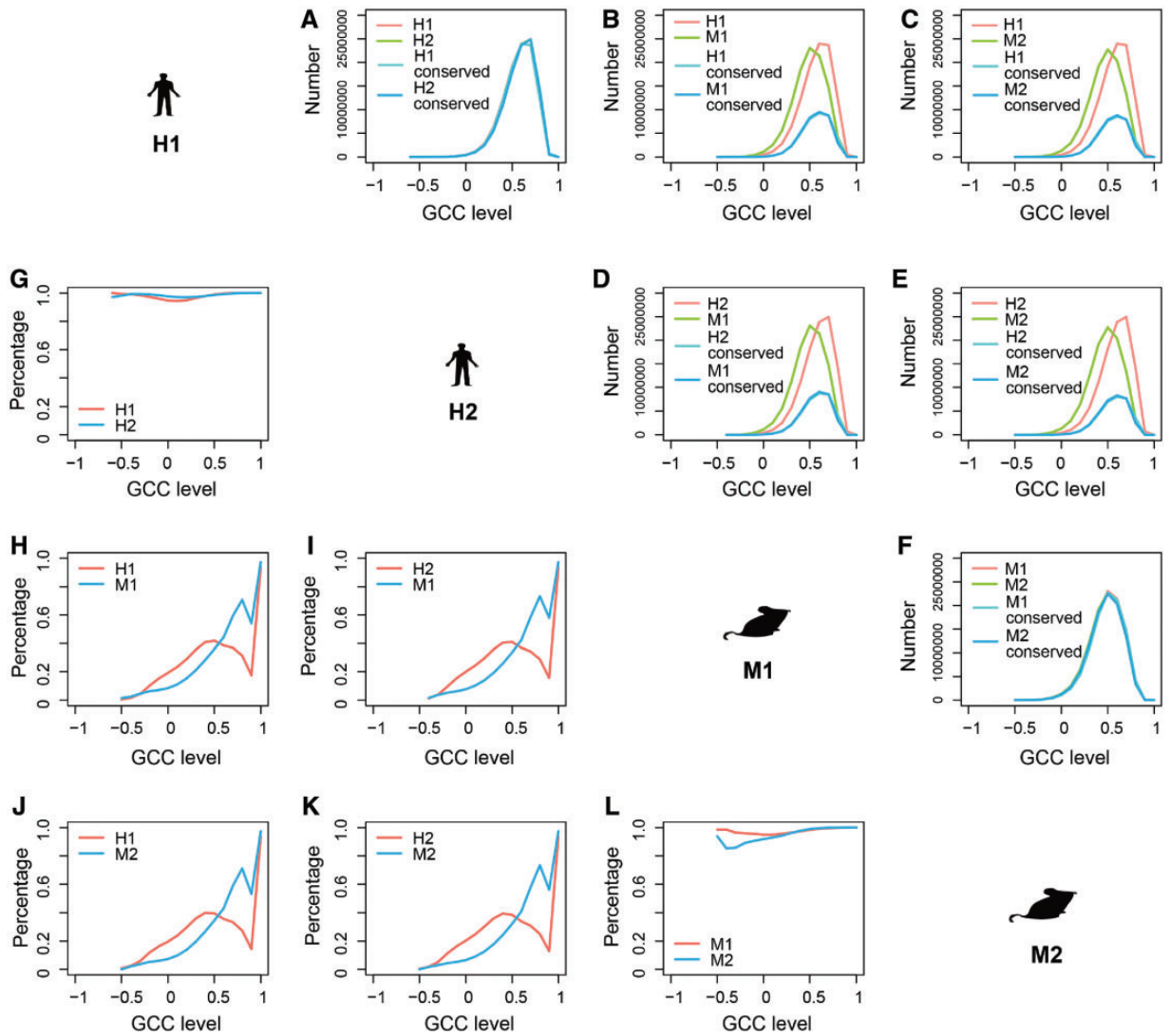

Figure 4. The distribution and the percentage of the conserved GCC. (A) The distribution of the conserved GCC in H1 and H2; (B) the distribution of the conserved GCC in H1 and M1; (C) the distribution of the conserved GCC in H1 and M2; (D) the distribution of the conserved GCC in H2 and M1; (E) the distribution of the conserved GCC in H2 and M2; (F) the distribution of the conserved GCC in M1 and M2. (G) the percentage of the conserved GCC in H1 and H2; (H) the percentage of the conserved GCC in H1 and M1; (I) the percentage of the conserved GCC in H2 and M1; (J) the percentage of the conserved GCC in H1 and M2; (K) the percentage of the conserved GCC in H2 and M2; (L) the percentage of the conserved GCC in M1 and M2.

The human–mouse general co-expression differences database (coexpressMAP)

To help researchers access and use the information of human–mouse general co-expression differences, we developed the coexpressMAP database that includes human gene GCC values, mouse gene GCC values and the differences between human and mouse. The coexpressMAP database is available at http://www.bioapp.org/coexpressMAP. Users can query the database by inputting two human (or mouse) genes. For example, if we input the two genes *CYP2F1* and *TNR* in human (or *Cyp2f1* and *Tnr* in mouse), the database will display the gene symbols, chromosome numbers, the GCC value in human (GCC = 0.846), the GCC value in mouse (GCC = −0.155) and the difference of GCCs between human and mouse (1.001). Comparison with the reference range of conserved co-expression relationships (−0.0795, 0.0795) reveals that this general co-expression difference is larger. That is, the gene pair *CYP2F1 (Cyp2f1)-TNR (Tnr)* had different co-expression relationships between human and mouse. The gene pair is labeled 'differenced' in the last column of the research results. As another example, if we input the gene pair *MYO1F (Myo1f)-HS6ST1 (Hs6st1)*, the database shows that the GCC value in human is 0.400, the GCC value in mouse is 0.398 and the difference of GCC between human and mouse is 0.002. This value of difference of GCC falls into the reference range of conserved co-expression relationships (−0.0795, 0.0795), and thus, the gene pair is labeled 'conserved' in the last column of the research results.

In addition, users can also download the human general co-expression maps (all the GCC values of human gene pairs), the mouse general co-expression maps (all the GCC values of mouse gene pairs), the list of conserved co-expression relationships between human and mouse, the HGCN and the MGCN from the link http://www.bioapp.org/coexpressMAP. Researchers can use these data to perform their own analyses.

Discussion

The mouse has a high similarity of sequence to human and is commonly used in laboratory study. However, similarity of

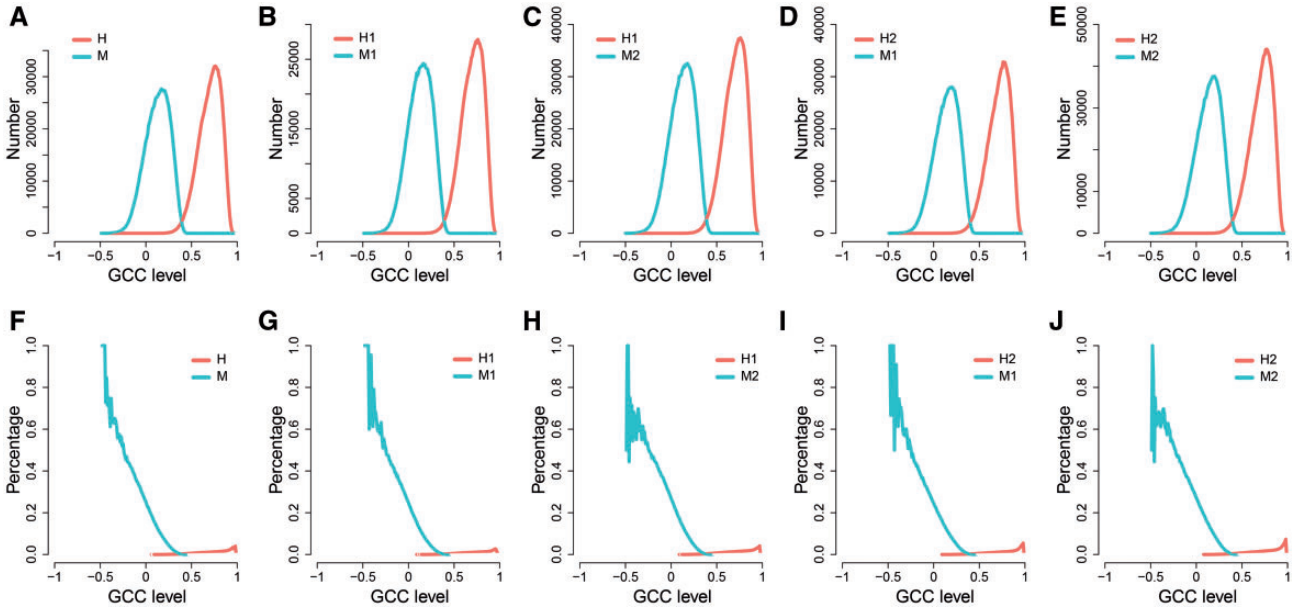

Figure 5. The distribution and the percentage of largely changed GCC (gene pairs showed lower GCC in mouse, but higher GCC in human). (**A**) The distribution of largely changed GCC in H and M; (**B**) the distribution of largely changed GCC in H1 and M1; (**C**) the distribution of largely changed GCC in H1 and M2; (**D**) the distribution of largely changed GCC in H2 and M1; (**E**) the distribution of largely changed GCC in H2 and M2; (**F**) the percentage of largely changed GCC in H and M; (**G**) the percentage of largely changed GCC in H1 and M1; (**H**) the percentage of largely changed GCC in H1 and M2; (**I**) the percentage of largely changed GCC in H2 and M1; (**J**) the percentage of largely changed GCC in H2 and M2.

genes at the sequence level does not immediately indicate that the functions of the genes are similar. In this study, we compared the general co-expression landscapes using human–mouse one-to-one orthologous genes and observed many obvious differences between the two species. (1) The general co-expression relationships had different distributions between human and mouse, such as in skewness, kurtosis, mean and median. As a whole, human genes had stronger co-expression relationships than mouse. (2) It is noteworthy that only about 36.54% of the general co-expression relationships were conserved between human and mouse. However, the average percentage of conserved general co-expression relationships is 98.58% in the same species (Table 2). This indicated that one-to-one orthologous gene pairs have a different ability of collaboration between human and mouse. (3) We further analyzed largely changed general co-expression relationships from mouse to human. We found that, for a gene pair, the higher the degree of negative correlation, the more likely it is to have a large change in the co-expression relationship. (4) In the aspect of general co-expression network, for the human–mouse one-to-one orthologous genes, we observed almost no correlation between different species on the four network topology characteristics (degree, betweenness, clustering coefficient and shortest path length), but a strong correlation within the same species (Figure 6, Supplementary Figures S3–S5). This also suggests that there is a large difference between human and mouse.

The entire analyses were repeated using another two independent sample sets, H2 (1000 human samples) and M2 (1000 mouse samples). These results still support the high differences of general co-expression relationships between the two species. For instance, we observed that the skewness of H1 (−0.813) was larger than that of M1 (−0.589), and this phenomenon was the same as H2 and M2 (−0.848 for H2 and −0.575 for M2; Figure 2B). A similar phenomenon was observed for kurtosis, mean, and median (Figure 2B; for detailed information, see Table 1). The percentages of conserved general co-expression relationships were consistent: 37.32% for H1-M1, 34.85% for H1-M2, 35.59% for H2-M1 and 33.02% for H2-M2. All the differences of the general co-expression network of HGCN1 and MGCN1 could be repeated with HGCN2 and MGCN2. Hence, our research is consistent, stable and repeatable.

In this study, we also compared the differences of network functional modules between human and mouse. First, we identified the modules for two HGCNs (HGCN1 and HGCN2) and two MGCNs (MGCN1 and MGCN2) using Cytoscape ClusterONE packages, which builds on the concept of the cohesiveness score and uses a greedy growth process to find clusters in a interaction network [15]. A module was identified by ClusterONE indicated that a group of genes in it has a higher general co-expression relationship each other. For modules that contain at least 100 genes, we then analyzed the intersection of modules between any two networks. The sorted gene numbers of intersection are drawn in Figure 7. We observed that the gene numbers of intersection within species were much greater than that between the two species. This indicated that the function modules still showed great differences between human and mouse. The modules for HGCN1, HGCN2, MGCN1 and MGCN2 can be found in Supplementary File S1 at http://www.bioapp.org/coexpressMAP/#Supplementary.

We further analyzed the functional similarities and differences of genes with high-degree between human and mouse. We first calculated and sorted the degrees of nodes for two general co-expression networks HGCN and MGCN. For the top 50% genes with high-degree, we divided these genes into 10 parts (top 0–5%, 5–10%, 10–15% and so on) and then annotated each part to gene ontology(GO) categories using DAVID [16]. Figure 8 shows the percentage of shared GO categories in human and mouse. We observed a decrease in the percentage of shared GO categories with a decrease in degree. This suggests that the more important the genes in the HGCN and MGCN network, the more functions they share between the two species. However, for genes with high betweenness, we did not observe the

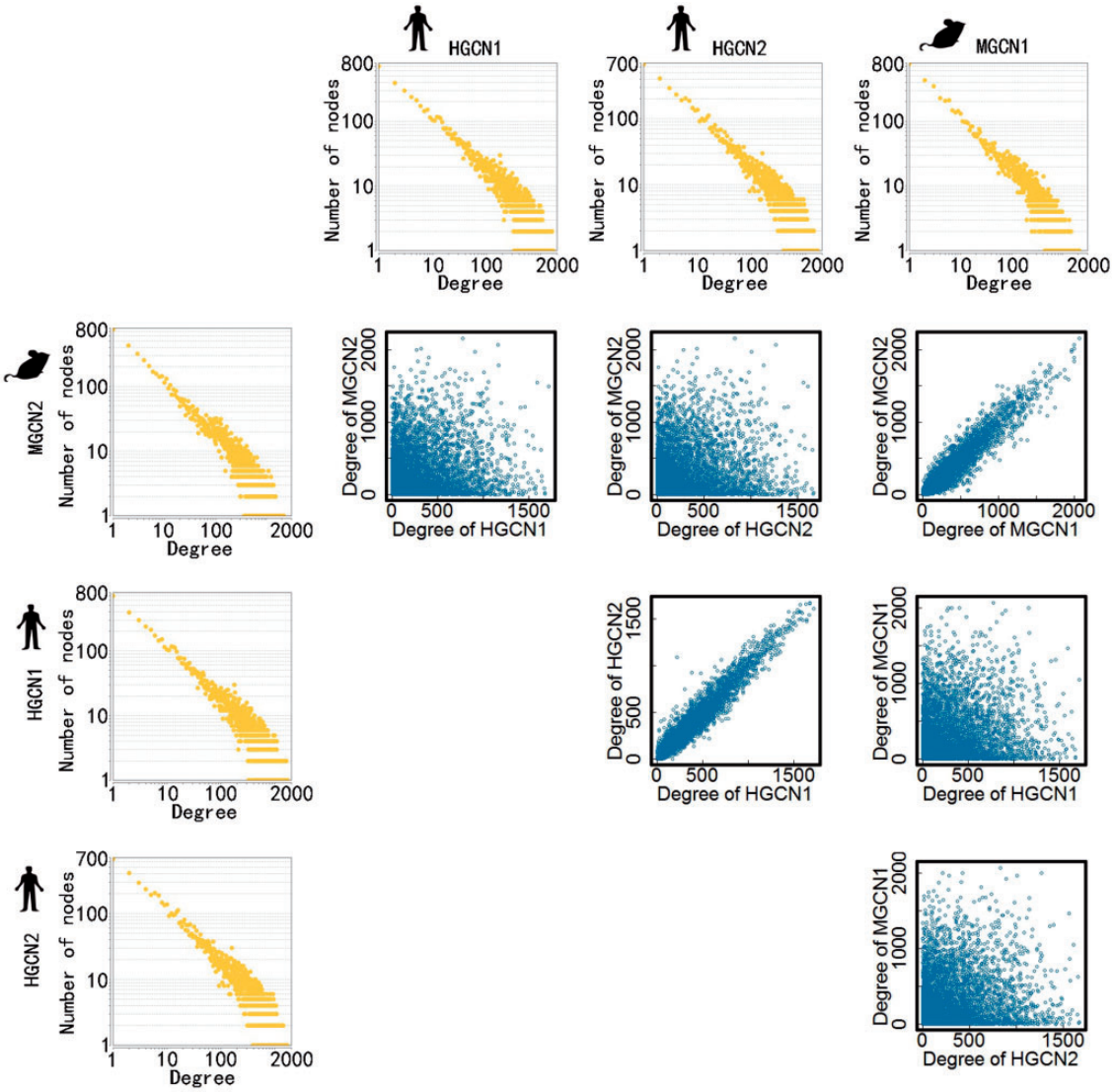

Figure 6. Comparison of the distribution and correlation of degrees between human and mouse.

correlations between top genes and the percentage of shared GO categories (Supplementary Figure S6). There still were larger functional differences between human and mouse. For the top 5% genes with a high degree, 61 GO categories were annotated in human, 132 GO categories were annotated in mouse and 40 GO categories were shared in both human and mouse. There were 21 human-specific GO categories (such as GO:0000045 autophagosome assembly) and 92 mouse-specific GO categories (such as GO:0000118 histone deacetylase complex). All the GO annotation results for the top 5–50% genes with a high degree can be found in Supplementary File S2 at http://www.bioapp.org/coexpressMAP/#Supplementary.

We also analyzed the gene substitutions for modules in GO functional categories. For HGCN and MGCN, we identified 13 human modules and 12 mouse modules (these modules can be found in Supplementary File S3 at http://www.bioapp.org/coexpressMAP/#Supplementary) that contain at least 100 genes using Cytoscape ClusterONE packages. We then annotated these modules to 910 GO categories. For each GO functional category, we analyzed gene substitutions for modules. We noticed that in positive regulation of target of rapamycin signaling

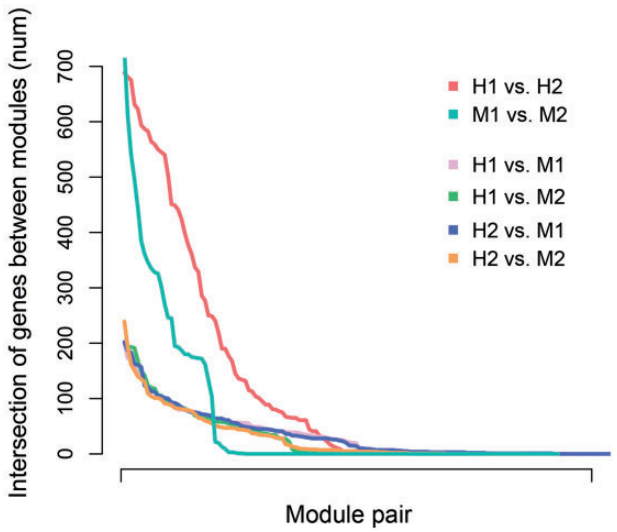

Figure 7. The intersection of genes between human and mouse modules.

category (GO: 0032008), five genes (*WDR59, RPTOR, WDR24, MLST8 and RHEB*) in human Module 6 and five genes (*Wdr59, Rptor, Wdr24, Mlst8 and GOLPH3*) in mouse Module 2 were annotated in the category. We can see the *GOLPH3* gene in the functional module in mouse but not in human, and the *GOLPH3* gene was replaced by the *RHEB* gene in human. Furthermore, we searched the published literatures for the function of five genes in the human module and found a strong correlation between the five genes. *WDR24* and *WDR59* are involved in the formation of a complex *GATOR2*, which is associated with the positive regulation of the *mTORC1* pathway [17]. Meanwhile, *mTORC1* is one of the two complexes of mechanistic target of rapamycin (*mTOR*) and is the hetero-oligomeric assembly of *mTOR*, *RAPTOR* (alias of *RPTOR*) and *MLST8*. *RHEB* can activate *mTORC1* as the GTP-bound form, and negatively regulate *mTORC1* through *TSC1* and *TSC2* [18, 19]. However, for *GOLPH3*, we do not found any direct interaction between *GOLPH3* and any one of the other four genes (*WDR24, WDR59, MLST8* and *RPTOR*). For the other 909 GO categories, 252 categories contain both human modules and mouse modules. The human-specific genes and mouse-specific genes in these modules can be found in Supplementary File S3.

In a word, after comparing the general co expression landscapes, we found significant differences between human and mouse, and only 36.54% of the co-expression relationships were conserved between human and mouse.

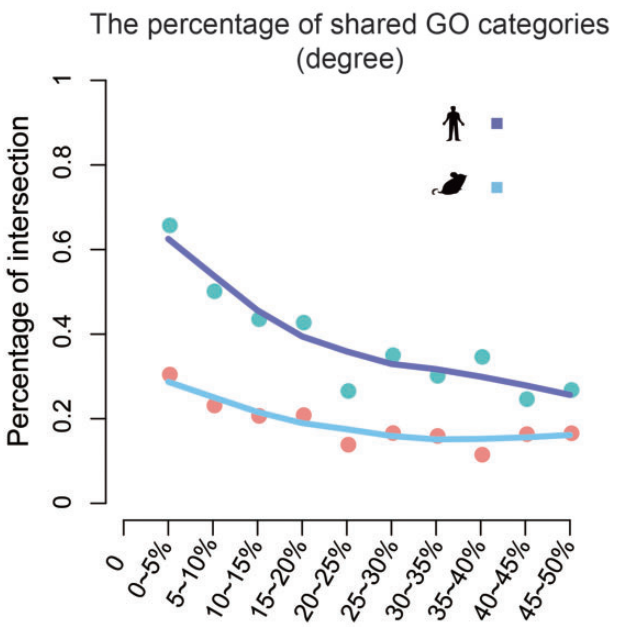

Figure 8. The percentage of shared GO categories between human and mouse for genes with high degree. We observed a decrease in the percentage of shared GO categories with a decrease in degree.

Methods

Human gene expression data

The human gene expression data were obtained from the NCBI GEO database. There were 122 474 public samples in the GPL570 platform (Affymetrix Human Genome U133 Plus 2.0 Array, including 54 675 probes) before 22 July 2016. We first searched the platform with the keywords 'control', 'healthy' and 'normal', and obtained 6901 samples. We then manually reviewed the sample information to confirm the normal samples. Finally, 3671 CEL files of normal human samples were downloaded, and the gene expression value was identified using the R 'affy' package [20].

Mouse gene expression data

The mouse gene expression data were also derived from the NCBI GEO database [21, 22]. The platform is GPL1261 (Affymetrix Mouse Genome 430 2.0 Array, including 45 101 probes). There were 48 713 public samples in the platform before 22 July 2016. A total of 5811 samples were obtained using the keywords 'wild type', 'wt', 'control', 'healthy' and 'normal'. We also manually reviewed the information of the 5811 samples to confirm the normal samples. Finally, 2361 normal mouse samples were used to identify the gene expression value. The CEL files of these samples were also analyzed using the R 'affy' package.

Human–mouse orthologous genes

The latest human–mouse orthologous gene list was downloaded from the Ensembl database mart 85 (ftp://ftp.ensembl.org/pub/release-85/mysql/ensembl_mart_85) [23–25]. In this study, the one-to-one orthologous genes were used to compare the general co-expression landscapes between human and mouse. We mapped the probes in the GPL570 platform and GPL1261 platform to these human and mouse genes based on their respective annotation files (downloaded from the GEO database). Finally, we built an integrated human–mouse gene expression matrix. The matrix consists of 14 331 rows (each row represents a gene, in total 14 331 human–mouse orthologous gene pairs) and 6032 columns (3671 normal human samples and 2361 normal mouse samples). Each human–mouse orthologous gene pair was assigned a fixed identification number (HgMg_id#). Using this approach, the general co-expression relationships can be one-to-one mapped between human and mouse.

The general co-expression coefficient

The general co-expression relationship was defined as the correlation between genes without considering cell or tissue type, cycle state or developmental stage between two genes in a species. We defined the GCC as:

$$GCC(i,j) = \frac{1}{n-1}\sum_{k=1}^{n}\left(\frac{g_{i,k}-\bar{g}_i}{S_{g_i}}\right)\left(\frac{g_{j,k}-\bar{g}_j}{S_{g_j}}\right).$$

Where $GCC(i,j)$ is the GCC between gene i and gene j, n is the number of the samples, $g_{i,k}$ is the expression value of gene i of sample k, $g_{j,k}$ is the expression value of gene j of sample k, $\bar{g}_i$ (or $\bar{g}_j$) is the mean of gene expression value of gene i (gene j) and S_{g_i} (or S_{g_j}) is the SD of gene expression value of gene i (gene j). GCC will reflect the collaboration between two genes in a species.

The general co-expression networks

We constructed two networks: the HGCN and the MGCN using the top 1 000 000 general co-expression relationships. Four network topology characteristics, degree, betweenness, clustering coefficient and shortest path length, were used to compare the differences and similarities between the two species [26–31]. The degree of the network is defined as the number of the edge connected with the node. The node betweenness is defined as the percentage of the number of the entire shortest path

passing through the node of the total shortest path. The clustering coefficient is the coefficient of the node aggregation in a graph. The shortest path of the two nodes is defined as the shortest path length. All the topology characteristics of networks were calculated using the software Cytoscape [32].

Key Points

- We compared the general co-expression maps between the human and mouse.
- We observed there exist low similarities of the co-expression relationships between the human and mouse.
- We provide the coexpressMAP database to search the co-expression relationships between human and mouse.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Funding

This work was supported in part by grants from the Natural Science Foundation of Heilongjiang Province (grant number C2016036). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

1. Cheon DJ, Orsulic S. Mouse models of cancer. *Annu Rev Pathol* 2011;**6**:95–119.
2. Sakamoto K, Schmidt JW, Wagner KU. Mouse models of breast cancer. *Methods Mol Biol* 2015;**1267**:47–71.
3. Dine J, Deng CX. Mouse models of BRCA1 and their application to breast cancer research. *Cancer Metastasis Rev* 2013;**32**:25–37.
4. Baribault H. Mouse models of type 2 diabetes mellitus in drug discovery. *Methods Mol Biol* 2016;**1438**:153–75.
5. Valkenburg KC, Pienta KJ. Drug discovery in prostate cancer mouse models. *Expert Opin Drug Discov* 2015;**10**:1011–24.
6. William Yang X, Gray M. Mouse models for validating preclinical candidates for Huntington's disease. In: DC Lo, RE Hughes (eds). *Neurobiology of Huntington's Disease: Applications to Drug Discovery*. CRC Press, Boca Raton, FL, 2011.
7. Mouse Genome Sequencing Consortium; Waterston RH, Lindblad-Toh K, *et al*. Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002;**420**:520–62.
8. Lin S, Lin Y, Nery JR, *et al*. Comparison of the transcriptional landscapes between human and mouse tissues. *Proc Natl Acad Sci USA* 2014;**111**:17224–9.
9. Tsaparas P, Marino-Ramirez L, Bodenreider O, *et al*. Global similarity and local divergence in human and mouse gene co-expression networks. *BMC Evol Biol* 2006;**6**:70.
10. Monaco G, van Dam S, Casal Novo Ribeiro JL, *et al*. A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue, pathway and disease levels. *BMC Evol Biol* 2015;**15**:259.
11. Yao P, Lin P, Gokoolparsadh A, *et al*. Coexpression networks identify brain region-specific enhancer RNAs in the human brain. *Nat Neurosci* 2015;**18**:1168–74.
12. Voss RH, Thomas S, Pfirschke C, *et al*. Coexpression of the T-cell receptor constant alpha domain triggers tumor reactivity of single-chain TCR-transduced human T cells. *Blood* 2010;**115**:5154–63.
13. Menashe I, Grange P, Larsen EC, *et al*. Co-expression profiling of autism genes in the mouse brain. *PLoS Comput Biol* 2013;**9**:e1003128.
14. Liu T, Yu L, Ding G, *et al*. Gene coexpression and evolutionary conservation analysis of the human preimplantation embryos. *Biomed Res Int* 2015;**2015**:316735.
15. Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods* 2012;**9**:471–2.
16. Dennis G, Jr, Sherman BT, Hosack DA, *et al*. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol* 2003;**4**:P3.
17. Chantranupong L, Scaria SM, Saxton RA, *et al*. The CASTOR proteins are arginine sensors for the mTORC1 pathway. *Cell* 2016;**165**:153–64.
18. Nakashima A, Kawanishi I, Eguchi S, *et al*. Association of CAD, a multifunctional protein involved in pyrimidine synthesis, with mLST8, a component of the mTOR complexes. *J Biomed Sci* 2013;**20**:24.
19. Oshiro N, Rapley J, Avruch J. Amino acids activate mammalian target of rapamycin (mTOR) complex 1 without changing Rag GTPase guanyl nucleotide charging. *J Biol Chem* 2014;**289**:2658–74.
20. Gautier L, Cope L, Bolstad BM, *et al*. affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 2004;**20**:307–15.
21. Edgar R, Barrett T. NCBI GEO standards and services for microarray data. *Nat Biotechnol* 2006;**24**:1471–2.
22. Barrett T, Wilhite SE, Ledoux P, *et al*. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res* 2013;**41**:D991–5.
23. Hubbard T, Barker D, Birney E, *et al*. The Ensembl genome database project. *Nucleic Acids Res* 2002;**30**:38–41.
24. Flicek P, Amode MR, Barrell D, *et al*. Ensembl 2012. *Nucleic Acids Res* 2012;**40**:D84–90.
25. Flicek P, Amode MR, Barrell D, *et al*. Ensembl 2014. *Nucleic Acids Res* 2014;**42**:D749–55.
26. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;**5**:101–13.
27. Grennan KS, Chen C, Gershon ES, *et al*. Molecular network analysis enhances understanding of the biology of mental disorders. *Bioessays* 2014;**36**:606–16.
28. Jiang X, Liu B, Jiang J, *et al*. Modularity in the genetic disease-phenotype network. *FEBS Lett* 2008;**582**:2549–54.
29. Ravasz E, Barabasi AL. Hierarchical organization in complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 2003;**67**:026112.
30. Park J, Barabasi AL. Distribution of node characteristics in complex networks. *Proc Natl Acad Sci USA* 2007;**104**:17916–20.
31. Zhang X, Zhang R, Jiang Y, *et al*. The expanded human disease network combining protein-protein interaction information. *Eur J Hum Genet* 2011;**19**:783–8.
32. Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.