SCIENCE CHINA
Life Sciences

THEMATIC ISSUE: Evolution of genes and genomes
•RESEARCH PAPER•

April 2019 Vol.62 No.4: 594–608
https://doi.org/10.1007/s11427-019-9483-6

Topological evolution of coexpression networks by new gene integration maintains the hierarchical and modular structures in human ancestors

Jian Zu[1,2*], Yuexi Gu[1], Yu Li[1], Chentong Li[1], Wenyu Zhang[3], Yong E. Zhang[4], UnJin Lee[2], Li Zhang[2] & Manyuan Long[2*]

[1]*School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China;*
[2]*Department of Ecology and Evolution, The University of Chicago, Chicago, Illinois, 60637, USA;*
[3]*Center for Systems Biology, Soochow University, Suzhou 215006, China;*
[4]*State Key Laboratory of Integrated Management of Pest Insects and Rodents & Key Laboratory of the Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China*

Received October 15, 2018; accepted November 5, 2018; published online March 21, 2019

We analyze the global structure and evolution of human gene coexpression networks driven by new gene integration. When the Pearson correlation coefficient is greater than or equal to 0.5, we find that the coexpression network consists of 334 small components and one "giant" connected subnet comprising of 6317 interacting genes. This network shows the properties of power-law degree distribution and small-world. The average clustering coefficient of younger genes is larger than that of the elderly genes (0.6685 vs. 0.5762). Particularly, we find that the younger genes with a larger degree also show a property of hierarchical architecture. The younger genes play an important role in the overall pivotability of the network and this network contains few redundant duplicate genes. Moreover, we find that gene duplication and orphan genes are two dominant evolutionary forces in shaping this network. Both the duplicate genes and orphan genes develop new links through a "rich-gets-richer" mechanism. With the gradual integration of new genes into the ancestral network, most of the topological structure features of the network would gradually increase. However, the exponent of degree distribution and modularity coefficient of the whole network do not change significantly, which implies that the evolution of coexpression networks maintains the hierarchical and modular structures in human ancestors.

Network biology, gene network evolution, scale-free network, natural selection, gene expression, self-organization, gene duplication

Citation: Zu, J., Gu, Y., Li, Y., Li, C., Zhang, W., Zhang, Y.E., Lee, U.J., Zhang, L., and Long, M. (2019). Topological evolution of coexpression networks by new gene integration maintains the hierarchical and modular structures in human ancestors. Sci China Life Sci 62, 594–608. https://doi.org/10.1007/s11427-019-9483-6

INTRODUCTION

Gene coexpression networks are of biological interest because coexpressed genes are controlled by the same transcriptional regulatory program, functionally related, or members of the same pathway or gene complex. Understanding the topological structure and evolutionary mechanisms of a human gene coexpression network is a fundamental step toward identifying the cell's internal organization and evolution (Barabási and Oltvai, 2004; Chen et al., 2013). In particular, the study of global structure of a human gene coexpression network contributes to revealing the function and behavior of a living cell, which is also va-

*Corresponding authors (Jian Zu, email: jianzu@xjtu.edu.cn; Manyuan Long, email: mlong@uchicago.edu)

luable for identifying gene functions or functional modules in metabolic pathways and signaling pathways. Moreover, understanding the basic organizing principles of human gene coexpression network is useful for identifying the function of younger genes and molecular mechanisms underlying the connections between genotype and phenotype (Barabási and Oltvai, 2004; Obayashi et al., 2012; Chen et al., 2013).

Due to the development of high-throughput data-collection techniques and new technology platforms, in the past few years, many theoretical and computational studies have led to significant conceptual advances regarding the global structure and evolution of gene or protein networks (e.g., Wagner, 2001, 2003; Maslov and Sneppen, 2002; Ravasz et al., 2002; Pastor-Satorras et al., 2003; Jordan et al., 2004; Horvath and Dong, 2008; Zhang et al., 2009; Yu et al., 2014; Zhang et al., 2015). A key aim of postgenomic biomedical research is to systematically catalogue all molecules and their interactions within a living cell. Because there is a clear need to understand how these molecules and the interactions between them determine the function of this enormously complex machinery, both in isolation and when surrounded by other cells (Barabási and Oltvai, 2004).

Previous studies on gene or protein networks indicate that the network exhibits a small-world pattern, very different from the properties displayed by purely random graphs (Barabási and Albert, 1999; Barabási and Oltvai, 2004). Within the cell, this ultra-small world effect was first documented for metabolism, where paths of only three to four reactions can link most pairs of metabolites (Jeong et al., 2000; Cohen and Havlin, 2003). This short path length indicates that local perturbations in metabolite concentrations could reach the whole network very quickly. Interestingly, the evolutionarily reduced metabolic network of a parasitic bacterium has the same mean path length as the highly developed network of a large multicellular organism, which indicates that there are evolutionary mechanisms that have maintained the average path length during evolution (Jeong et al., 2000; Cohen and Havlin, 2003; Barabási and Oltvai, 2004).

Moreover, it was found that the degree distribution of gene networks approximates a power law distribution with a well-defined cut-off, in which most of the nodes have only a few links, and only a few nodes with a very large number of links (Barabási and Albert, 1999; Barabási and Oltvai, 2004). The first evidence came from the analysis of metabolism, the analysis of the metabolic networks of 43 different organisms from all three domains of life (eukaryotes, bacteria, and archaea) indicates that the cellular metabolism has a scale-free topology, in which most metabolic substrates participate in only one or two reactions, but a few, such as pyruvate or coenzyme A, participate in dozens and function as metabolic hubs (Jeong et al., 2000; Ravasz et al., 2002). Further examples of scale-free organization include genetic regulatory networks, in which the nodes are individual genes and the links are derived from the expression correlations that are based on microarray data, or protein domain networks that are constructed on the basis of protein domain interactions (Jordan et al., 2004; Lee et al., 2004). The topology of a human gene coexpression network, derived from tissue-specific expression profiles, also shows scale-free property, which implies that the evolutionary self-organization of gene coexpression network via preferential node attachment. The ubiquity of scale-free network topological patterns suggests that network growth by preferential attachment, via mechanisms such as gene duplication, is a fundamental and conserved evolutionary process. However, not all networks within the cell are scale-free. For example, the transcription regulatory networks of *S. cerevisiae* and *Escherichia coli* offer an interesting example of mixed scale-free and exponential characteristics (Shen-Orr et al., 2002). In addition, the cellular networks that have been studied so far, including protein interaction and gene networks, have a high average clustering coefficient, which indicates that high clustering is a generic feature of biological networks (Clauset et al, 2004; Barabási and Oltvai, 2004).

In order to explore the evolutionary mechanism of human gene coexpression network, Jordan et al. (2004) proposed a model for how natural selection could influence gene expression divergence, mainly based on rapid, adaptation-driven divergence and convergent evolution of gene expression patterns. They found that genes with numerous coexpressed partners evolve more slowly on average than genes with fewer coexpressed partners, and genes that are coexpressed show similar rates of evolution. Thus, the strength of selective constraints on gene sequences is affected by the topology of the gene coexpression network. Zhang et al. (2015) examined the topological and functional evolution of a human gene coexpression network and found that over 5,000 new genes were integrated into gene-gene interaction networks throughout vertebrate evolution. These new genes experienced a gradual integration process into gene-gene interaction networks, starting on the network periphery and gradually acquiring increasing number of gene partners; some human-specific new genes finally evolved into hub structures with critical phenotypic effects.

Moreover, at the genomic level, gene duplication is thought to be a dominant evolutionary force in shaping biological networks (such as gene regulatory networks and protein-protein interaction networks). Chung et al. (2003) used combinatorial probabilistic methods to examine the evolution of graphs by node duplication processes. Both full duplication of nodes (with all their connections) as well as partial duplication (with only some connections) are analyzed. They found that partial duplication can produce power-law graphs with exponents less than 2, consistent with current data on biological networks. The power-law ex-

ponent for large graphs depends only on the growth process, not on the starting graph.

It has emerged that two fundamental processes have a key role in the development of real networks (Pastor-Satorras et al., 2003; Chung et al., 2006). First, most networks are the result of a growth process, during which new nodes join the system over an extended time period. Second, nodes prefer to connect to nodes that already have many links, a process that is known as preferential attachment. Indeed, if a node has many links, new nodes will tend to connect to it with a higher probability. This node will therefore gain new links at a higher rate than its less connected peers and will turn into a hub. Growth and preferential attachment have a common origin in protein networks that is probably rooted in gene duplication. In addition to natural selection, an emphasis has recently been placed on the role of fundamental physical principles in imposing order on biological systems (Wagner, 2003; Barabási and Oltvai, 2004).

Despite the significant advances in the past few years, molecular network biology is only in its infancy. Much remains unknown about the gene age-specific topological structure and evolution of gene or protein network. Subtle alterations in genetic interactions may lead to unexpectedly different regulatory dynamics and hence significant phenotypic changes (Wagner, 2003; Zhang et al., 2010, 2011). Particularly, there has been little discussion about how could natural selection have shaped the global structure of human gene coexpression network and how natural selection could influence gene expression divergence? During the evolutionary process of gene coexpression network, little is known about which topological structure features remain unchanged. In this study, our goal is to help understand the gene-age specific characteristics of human gene coexpression network and its evolutionary mechanisms. Based on the human gene coexpression database, we will analyze in quantifiable terms the topological properties of a human gene coexpression network and explore its formation mechanism. Particularly, based on the analysis results, we will propose a novel evolutionary model of gene coexpression network, which can help us construct a biomolecular network with comprehensive characteristics. Specifically, we emphasize the following four questions: (1) What are the global structure features of a human gene coexpression network (HGCN)? (2) What are the differences between the topological structure of elderly genes and that of younger genes? (3) How are younger genes integrated into the ancestral gene coexpression network and rewiring the gene coexpression network? (4) How do younger genes change the topological structure of ancestral gene coexpression network? In general, this study help us reveal the basic topological features of a gene coexpression network and provide some new insights for development of a novel evolutionary model of gene-gene interaction network.

RESULTS

The human gene coexpression network contains one "giant" connected subnet

The human gene correlation table is downloaded from http://coxpresdb.jp/download.shtml (COXPRESdb ver. 6.0) (Obayashi et al., 2012). In this study, the pairwise correlations between gene expression patterns were used to derive the gene coexpression network. When the Pearson correlation coefficient (PCC), $|r|\geqslant 0.5$ (that is, the correlation is relatively strong), we find that the human gene coexpression network (HGCN) has N=7,170 genes (nodes) and M=127,132 links (edges), where most of the pairwise genes are positive correlation and only nine pairwise genes are negative correlation (File S1 in Supporting Information, Figure 1). Specifically, this network consists of 334 small components involving only a few genes (A component or subnet of the network is a group of genes that interact with each other but do not interact with any other gene) and one "giant" connected subnet comprising of 6,317 interacting genes.

Moreover, based on the information of gene age data (File S2 in Supporting Information) (Zhang et al., 2010, 2011), we further find that 28.66% (2,055 genes) of genes in this network are younger genes (younger genes are genes which belong to branches 1–13), which has 32,035 links (accounting for 25.2% of all links) (Tables 1 and 2). For simplicity, the genes in branch 0 are called as elderly genes (5,115 genes, gene age group$\leqslant$–400), the genes in branch 1–3 are called as middle-aged genes (1,271 genes, gene age group –400 to –200), the genes in branch 4–8 are called as youth genes (630 genes, gene age group –200 to –50), the genes in branch 9–13 are called as juvenile genes (154 genes, gene age group –50 to 0).

The human gene coexpression network shows a power-law degree distribution

The average degree of a network tells us how many connections a gene has on average. By direct calculation, we find that the average degree of whole network and "giant" connected component is 36 and 40, respectively. The average degree of younger genes (branch 1–13) is less than that of the elderly genes (31 vs. 37) (Table 2). Particularly, with the increase of gene age, the average degree of genes would gradually increase (Figure 2A). Moreover, we find that 52.21% of younger genes (branch 1–13) would become a hub gene (a hub gene is the gene with degree $k\geqslant 6$). With the increase of gene age, the fraction of hub genes would also gradually increase (Figure 2B).

The degree distribution of a network $P(k)$, gives the probability that a selected gene has exactly k links. It is found that the degree distribution $P(k)$ of the whole network ap-

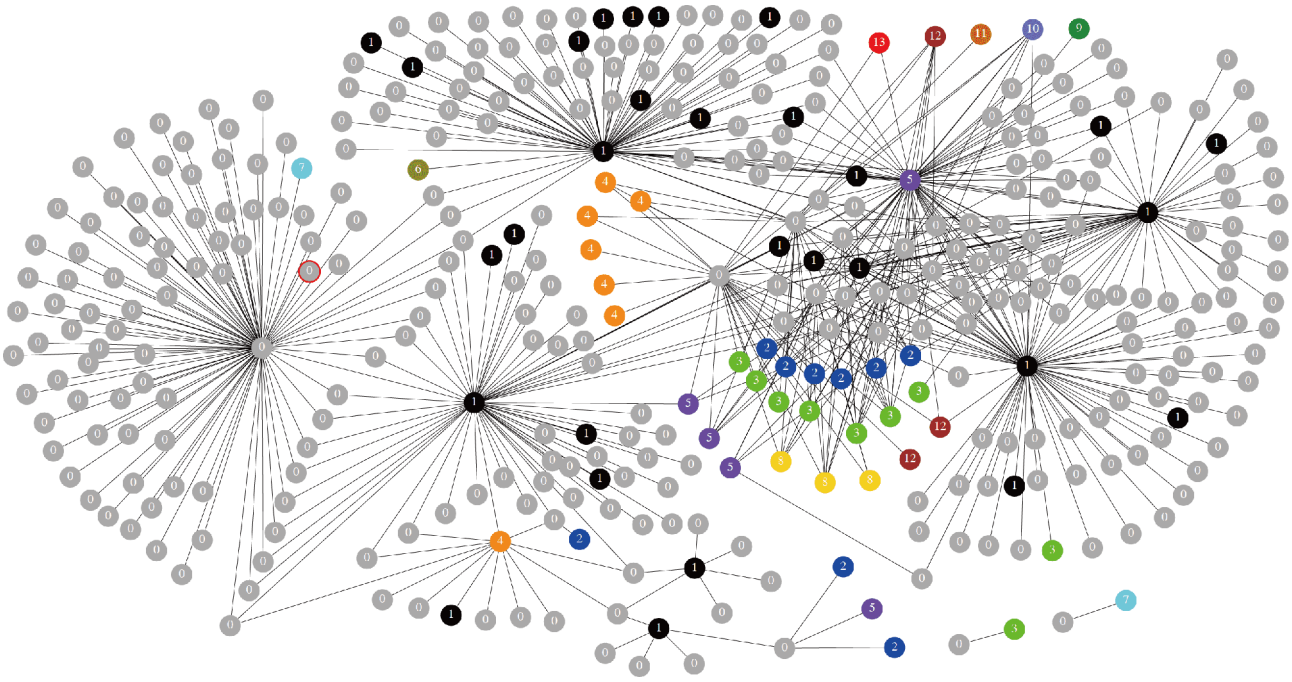

Figure 1 A part of the gene age-specific coexpression network in human. Each circle corresponds to a gene (node), and each link (edge) connecting two genes corresponds to a coexpression of the two genes. The color and number of a node indicate the gene age and branch number, respectively. The younger the gene is, the bigger the branch number it is assigned.

Table 1 Gene age and gene numbers of each branch[a)]

Branches	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Species	Zebrafish	*X. tropicalis*	Lizard	Platypus	Opposs-um	Elephant	Cow	Mouse	Marmoset	Rhesus	Gibbon	Orangu-tan	Chimpan-zee	Human
Divergence time (Myr)	–435.0	–352.0	–312.0	–177.0	–159.0	–105.0	–96.0	–91.9	–43.0	–29.3	–20.2	–15.8	–6.7	0.0
Gene age (Myr)	⩽–400.0	–393.5	–332.0	–244.5	–168.0	–132.0	–100.5	–93.95	–67.45	–36.15	–24.75	–18.00	–11.25	–3.35
# of genes	12,189	2,186	668	1,200	1,070	1,144	162	66	154	120	82	77	162	182
# of genes in HGCN	5,115	681	195	395	265	255	38	22	50	41	26	25	43	19
# of DNA-based duplicates in HGCN	—	492	115	219	152	142	21	16	29	15	9	14	19	4
# of RNA-based duplicates in HGCN	—	11	9	8	8	17	1	1	1	1	2	0	0	2
# of orphan genes in HGCN	—	51	16	42	13	13	1	0	2	0	0	0	1	3

a) The unit of time and age is million year (Myr), –3.35 Myr means 3.35 million years ago. Gene age is calculated as the middle time point of each branch. For example, genes assigned to branch 13 are shown at –3.35 Myr, the average origination time for an interval ranging from –6.7 to 0.0 Myr. The gene age of oldest branch (branch 0) is set as older than –400.0 Myr. HGCN denotes human gene coexpression network. # denotes number.

proximates a power-law distribution, that is, $P(k)=\alpha k^{-\beta}$, where α=0.318(95% CI, 0.256–0.395) and β=1.264(95% CI, 1.221–1.306) (Figure 2C and D). This implies that very many genes with only a few links, and a few genes with large numbers of links. Moreover, we find that the degree distribution of the biggest connected component also follows a power-law distribution, namely, $\hat{P}(k) = 0.333k^{-1.248}$ (Table 2). Interestingly, it is found that the degree distribution of younger genes (branch 1–13) also follows a power-law distribution (Table 2, Figure 2E), which indicates that many younger genes with only a few links and a few younger genes with large numbers of links.

The distribution of shortest path length approximates a Poisson distribution

The average shortest path length l, represents the average over the shortest paths between all pairs of genes and offers a measure of a network's overall navigability. We find that the average shortest path length of human gene coexpression network is 4.2; the maximal value of the shortest path length is 16. In particular, it is found that the average shortest path length l is proportional to the logarithm of network size N, that is to say, l=4.20~log(7170)=8.88, which indicates that the HGCN is characterized by the so called small-world property. Moreover, we can see that for both the whole co-

Table 2 Global statistical features of human gene coexpression network

Types	Whole network	Elderly genes (Branch 0)	Younger genes (Branch 1–13)	Giant connected component	Elderly genes (Branch 0)	Younger genes (Branch 1–13)
Number of genes	7,170	5,115	2,055	6,317	4,602	1,715
Average degree	36	37	31	40	41	37
Exponent of degree distribution (95% CI)	1.264 (1.221, 1.306)	1.193 (1.149, 1.236)	0.985 (0.924, 1.046)	1.248 (1.205, 1.291)	1.181 (1.138, 1.225)	0.952 (0.893, 1.010)
Average shortest path length	4.20	4.11	4.46	4.20	4.11	4.47
Average clustering coefficient	0.6019	0.5762	0.6685	0.5960	0.5720	0.6606
Average node betweenness	8911.0	9270.6	8016.0	10114.1	10303.9	9604.9
Average edge betweenness	659.5	658.8	660.5	664.8	662.5	667.9
Modularity	0.6622	—	—	0.6579	—	—
Number of components	335	—	—	1	—	—

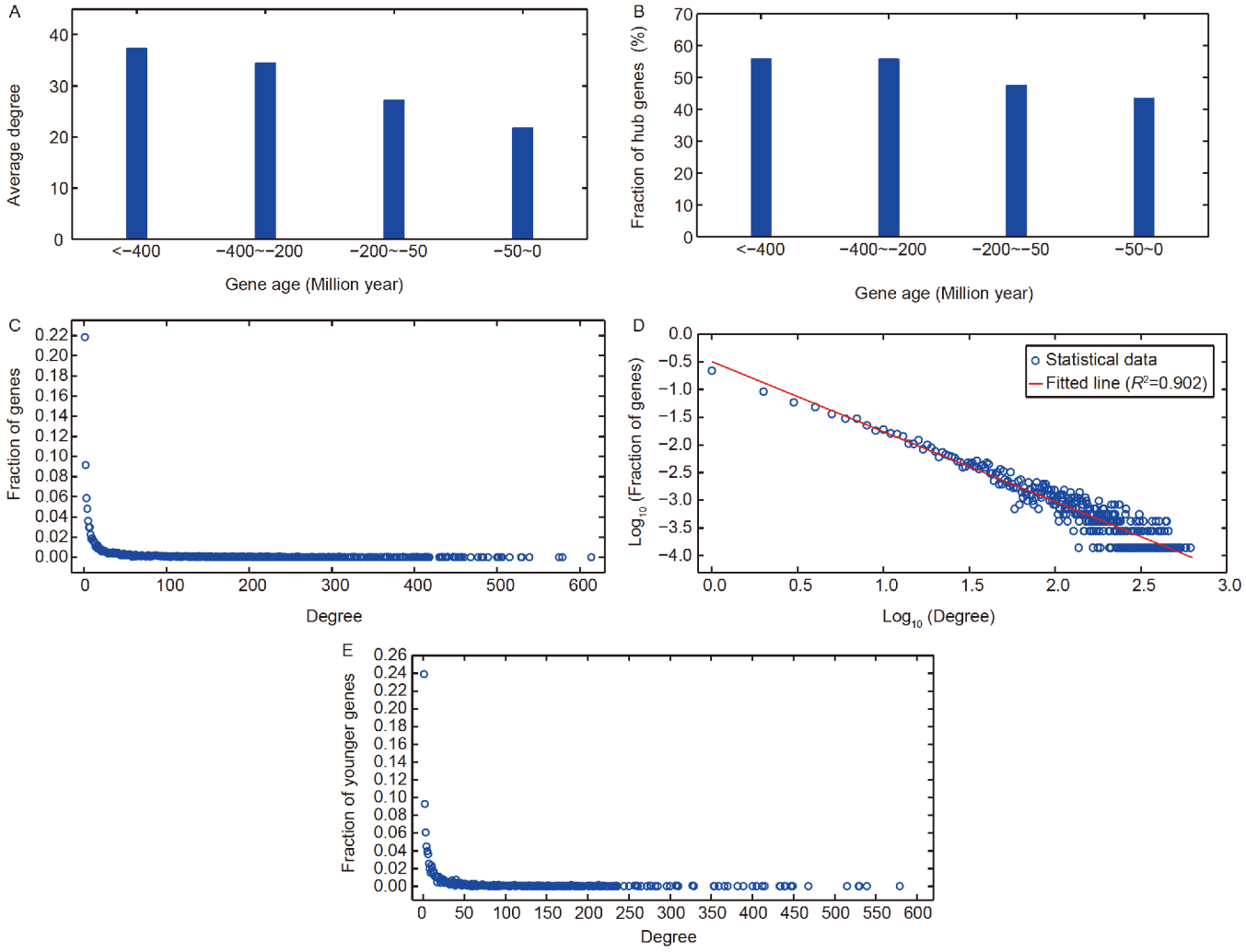

Figure 2 The degree and degree distribution of human gene coexpression network. A, The gene age-specific average degree. B, The gene age-specific proportion of hub genes: the number of hub genes with degree $k \geqslant 6$ accounts for the total number of genes in corresponding gene age group. C, The degree distribution of whole gene coexpression network. D, The log degree distribution of whole gene coexpression network. E, The degree distribution of younger genes (branch 1–13).

expression network and the "giant" connected component, the average shortest path length of younger genes is a little larger than that of the old genes (Table 2). With the increase of gene age, the average shortest path length of genes would gradually decrease (Figure 3A).

The distribution of shortest path length $D(l)$, gives the

probability that a selected pair of genes has exactly the shortest path length l. Specifically, we can see that the distribution $D(l)$ of shortest path length for whole network approximates a Poisson distribution, that is, $D(l)=\mu\lambda^{l}e^{-\lambda}/l!$, where μ=1.184(95% CI, 0.936–1.431 and λ=4.273(95% CI, 3.668–4.877) (Figure 3B). Interestingly, for both the whole network and the "giant" connected component, the shortest path length distribution of younger genes and elderly genes also follows a Poisson distribution (Figure 3C, Table 2), which implies that this network has a good navigability and the small-world property.

The gene coexpression network shows a hierarchical architecture

Average clustering coefficient characterizes the overall tendency of genes to form clusters or groups. The average clustering coefficient of whole network is 0.6019, which indicates that the genes of this network have a relatively strong tendency to form clusters or groups; the network shows a property of high clustering. Strikingly, for both the whole network and giant component, we can see that the average clustering coefficient of younger genes is larger than that of the old genes (Table 2). This implies that the younger the gene, the stronger the tendency to form clusters or groups. With the increase of gene age, the average clustering coefficient of genes would gradually decrease (Figure 4A).

The average clustering coefficient of all genes with k links, $C(k)$, is depicted in Figure 4B and C. We can see that when the degree of genes $k\geqslant 243$, then $C(k)=120.78k^{-0.944}$, the 95% confidence interval of exponent is (–1.031, –0.8562), which demonstrates that the genes with a larger degree show a property of hierarchical architecture. In other words, the communication between the different highly clustered neighborhoods is maintained by a few hub genes. Particularly, it is found that the younger genes with a larger degree also show a property of hierarchical architecture (Figure 4D).

Younger genes play an important role in the overall pivotability of network

The betweenness of a gene vs defined by the number of shortest paths going through the gene v. By direct calculation, we find that the average node betweenness of whole network is 8,911, the average node betweenness of youth genes (branch 4–8) is the biggest (Figure 5A). Moreover, for the whole network and giant component, the average node betweenness of younger genes is less than that of elderly genes (Table 2). But it is still very large, indicating that the younger genes also play an important role in the overall pivotability of the network. The removal of any of them may affect the communication between many pairs of genes through the shortest paths between them.

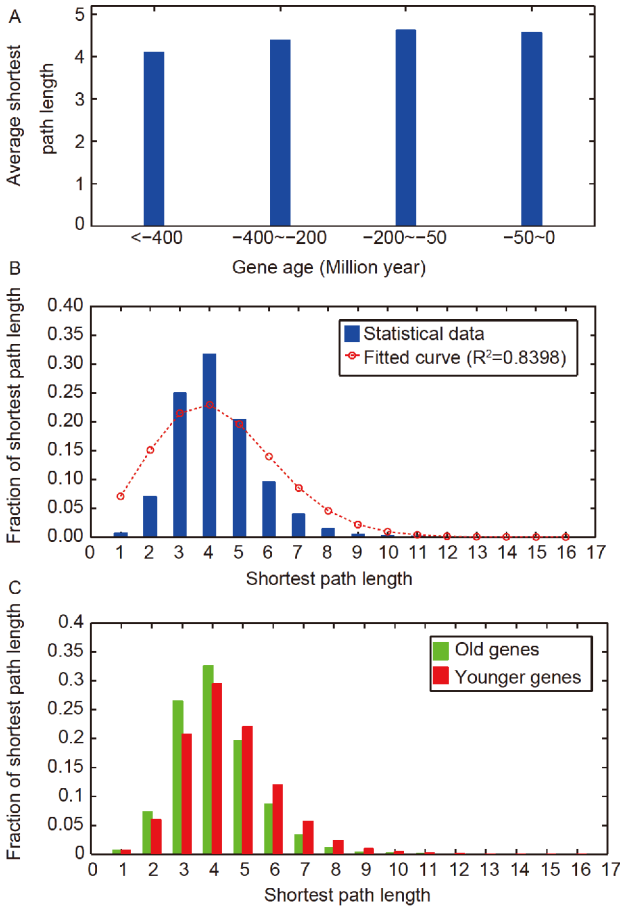

Figure 3 Average shortest path length and distribution. A, The gene age-specific average shortest path length. B, The distribution of shortest path length for whole network. C, The distribution of shortest path length for elderly genes and younger genes in the whole network.

On the other hand, the edge betweenness is defined as the number of the shortest paths that go through an edge (link) E in a network. We find that the average edge betweenness of whole network is 659.5, the average edge betweenness of youth genes (branch 4–8) is the biggest (Figure 5B). Particularly, for the whole network and giant component, the average edge betweenness of younger genes is a little larger than that of the old genes (Table 2), which further indicates that the younger genes and their new links play an important role in the overall pivotability of the network. A link with a younger gene added represents a bridge-like connector between two parts of a network, and the removal of which may significantly affect the communication between many pairs of genes through the shortest paths between them.

The network shows a significant community structure

The modularity of a network with respect to some division (or gene communities) measures how good the division is, or how separated are the different gene communities from each other, in the sense that there are many links within communities and only a few between them. If the fraction of

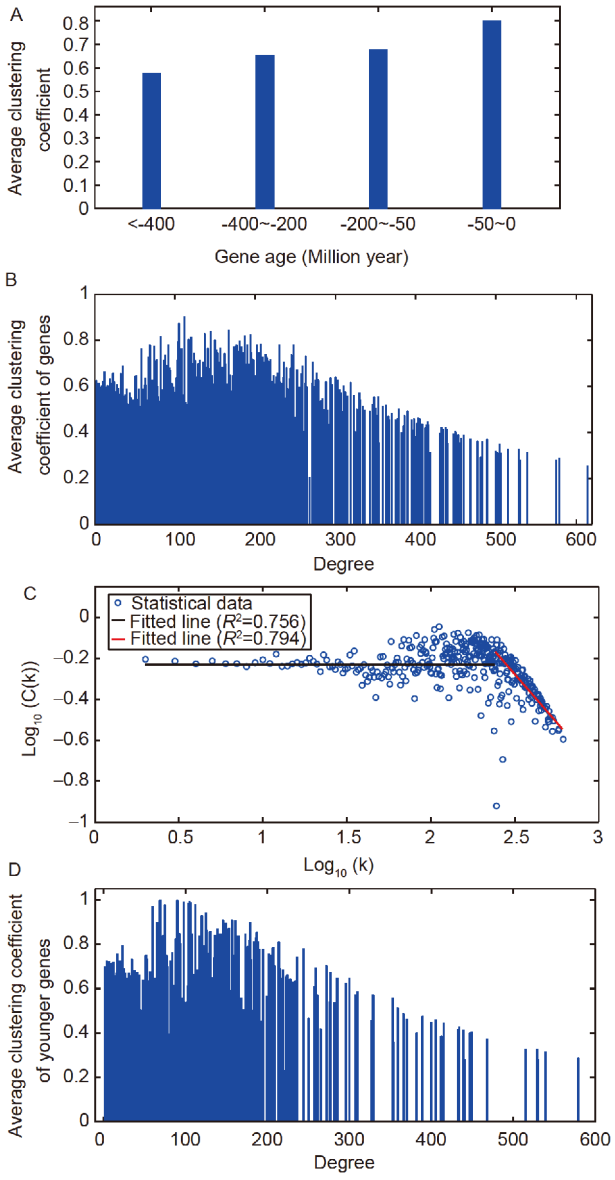

Figure 4 Average clustering coefficient and distribution. A, The gene age-specific average clustering coefficient. B, The average clustering coefficient of genes with degree *k*. C, The log average clustering coefficient of genes with degree *k*. D, The average clustering coefficient of younger genes with degree *k*.

within-community edges is no different from what we would expect for the randomized network, the modularity will be zero. We find that for both the whole network and giant component, the modularity coefficient is larger than 0.65 (Table 2), which implies that the network may have a significant community structure. In other words, there exists a good division of the network into small communities; different communities may show different functions.

The coexpression network contains few redundant duplicate genes

The studied HGCN contains 2,055 younger genes (branch 1–13). Among these younger genes, based on the updated data of gene origination mechanism (File S3 in Supporting Information) (Zhang et al., 2010, 2011), we find that at least 63.65% (1,308/2,055) of younger genes are created from duplication-based mechanisms (either from DNA-level duplication or RNA-level duplication), which is in line with the classic argument that gene duplication is the dominant force of evolution (Chung et al., 2003; Pastor-Satorras et al., 2003; Barabási and Oltvai, 2004). Therefore, another interesting question is about do most duplicate genes eventually lose all common interactions, or do they retain some common interactions indefinitely?

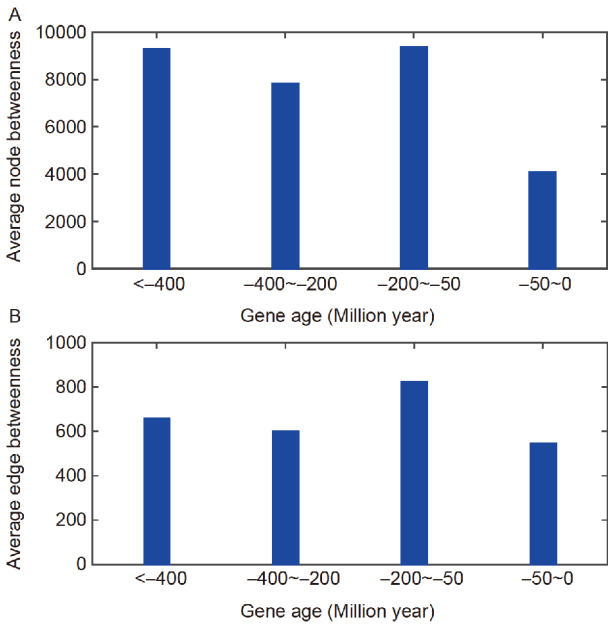

Figure 5 Average node betweenness and average edge betweenness. A, The gene age-specific average node betweenness of whole network. B, The gene age-specific average edge betweenness of whole network.

Eventually, we find that only 15.67% (205/1,308) of duplicate gene pairs shares an interaction partner. Among the 205 duplicate genes, on average, they only inherit 33.49% of linking partners from their parental genes (14 vs. 42). From Figure 6A, we can see that the younger the duplicate genes, the more the common shared interactions. If the number of common interaction partners is taken as a measure of functional overlap, we can say that the coexpression network contains few redundant duplicate genes. In addition, we find that only 14.3% (187/1,308) of duplicate genes finally shows the feature of interaction with their parental genes (parent-child interaction). The younger the duplicate genes, the more the parent-child interactions (Figure 6B).

Evolution of coexpression network through gene duplication

Based on the dataset of gene origination mechanism (File S2

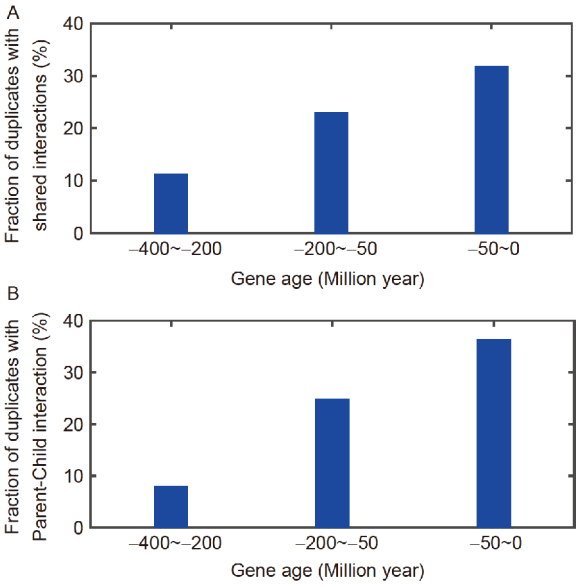

Figure 6 Fraction of duplicate genes with shared interactions. A, Histogram of the fraction of duplicate genes which share at least one common interacting gene with their parents. B, The fraction of duplicate genes which finally shows the feature of parent-child interaction.

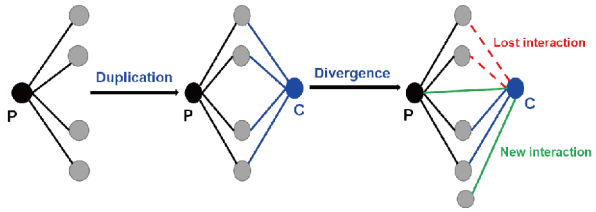

Figure 7 The evolutionary process of a gene duplication and divergence. Shortly after a gene duplication, the parental gene P and duplicate gene C will interact with the same genes. Eventually, some or all of the common interactions will be lost, and new interactions may be gained by the duplicate gene C. In the rightmost panel, gene C has lost two common interactions and gained two new interaction partners.

in Supporting Information), we find that among the 2,055 younger genes (branch 1–13), 60.68% (1,247/2,055) of them originate from DNA-based duplication, 2.97% (61/2,055) of them originate from RNA-based duplication and 6.91% (142/2,055) of them are orphan genes (Tautz and Domazet-Lošo, 2011). Moreover, the origination mechanism of other genes (605/2,055=29.44%) are ambiguous, some may be orphan genes, the other may originate from gene duplication or other mechanisms. Therefore, gene duplication is still considered as a dominant evolutionary force in shaping this coexpression network. The second evolutionary force is by creating an orphan gene.

Firstly, we consider how is a duplicate gene integrated into the ancestral gene coexpression networks? We assume that the initial gene coexpression network G_0 only consists of genes in branch 0. The genes in branch 1–13 are added into the initial network branch by branch. In particular, we assume that there is no interaction lost for parent genes and other existing genes. In this case, the evolutionary process of a gene duplication and divergence is depicted in Figure 7.

In this study, we analyze 19,462 genes in human, during the period of 390.15 million years (from branch 2 to branch 13) we find that there are at least 805 duplicate genes (755 genes are DNA-based duplication and 50 genes are RNA-based duplication), therefore the effective duplication rate of each gene exceeds $805/(393.5–3.35)/19462=1.06\times10^{-4}$/Myr. Moreover, it is found that more than 75% of parent genes have a degree less than 6, and the average degree of their parent genes is 13, which is much less than the average degree of whole network 36. Therefore, we can say that the gene duplication usually occurs randomly from the existing genes with a small degree. That is, there is a degree-biased gene duplication during the evolutionary process of coexpression network.

After a gene duplication, we find that the 805 duplicate genes (from branch 2 to branch 13) lose 8,500 links compared with the 9,859 links of their parent genes, therefore the rate of interaction loss for each duplication gene is $(8500/9859)\times(1/(393.5–3.35))=2.21\times10^{-3}$/Myr. Overall, about 86.22% of the common interactions inherited from their parent genes are lost for each duplication gene.

Moreover, among the 805 duplicate genes, there are 145 duplicate genes which have a link with their parent genes, that is, only 18.0% of duplicate genes show the property of parent-child interaction. On average, the rate of parent-child interaction for each duplication gene is $(145/805)\times(1/(393.5–3.35))=4.62\times10^{-4}$/Myr. In addition, except for the parent-child interaction and shared interactions with their parent genes, the 805 duplicate genes from branch 2 to branch 13 regain 20,541 new links, therefore the rate of new interaction for each duplication gene is $(20541/805)\times(1/(393.5–3.35))=6.54\times10^{-2}$/Myr. In general, the average new interaction of each duplication gene is 26.

Interestingly, we find that the average degree of new interaction partners is 161, and more than 80% of the new interaction genes have a degree larger than 35 (Figure 8A). Therefore, we can say that the new duplicate genes develop new links through a ‘rich-gets-richer’ mechanism: the more connected a gene is, the more likely it is that new duplicate genes will link to it, which allows the highly connected genes to acquire new links faster than their less connected peers. In other words, there exists a preferential principle for new duplicate genes to develop new links, the new duplicate gene prefers to link to an existing gene that is already highly connected (Barabási and Oltvai, 2004).

Evolution of coexpression network by creating orphan genes

Based on the dataset of gene origination mechanism (File S2

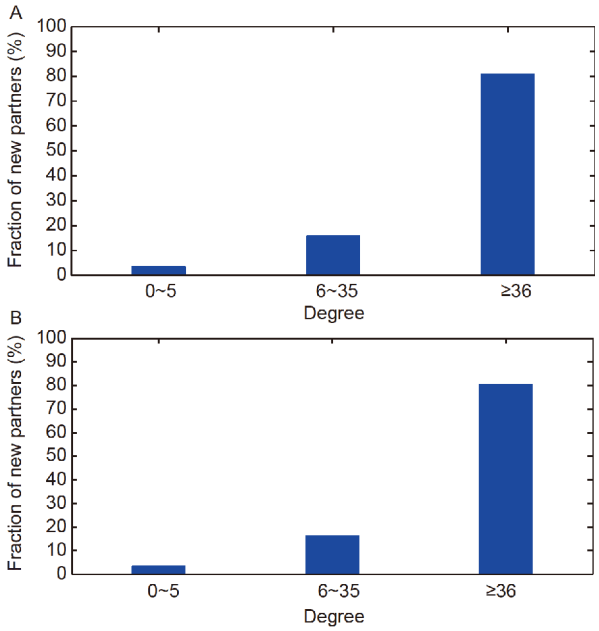

Figure 8 The fraction of new interaction partners with different degrees. A, The new interaction partners developed by a duplicate gene. B, The new interaction partners developed by an orphan gene.

in Supporting Information), we find that there are 91 orphan genes from branch 2 to branch 13. Therefore, the occurrence rate of an orphan gene exceeds $91/(393.5–3.35)=2.33\times10^{-1}$/Myr. On the whole, the 91 orphan genes from branch 2 to branch 13 gain 1,624 new links, so the rate of new interaction for each orphan gene is $(1624/91)\times(1/(393.5–3.35))=4.57\times10^{-2}$/Myr. In particular, the average degree of an orphan gene is 18, and the average degree of their new interaction partners is 133, more than 80% of the new interaction partners have a degree larger than 35 (Figure 8B). Therefore, we conclude that the attachment of an orphan gene also keeps a 'rich-gets-richer' mechanism.

Evolutionary model of human gene coexpression network

Based on the two types of evolutionary forces, we obtain the following evolutionary model of human gene coexpression network (Figure 9). Specifically, the evolutionary model is defined by the following rules. We start from a gene coexpression network G_0, which only consists of genes in branch 0, and each time step we perform the following operations: (i) Addition of genes: For one thing, biased gene duplication: the effective duplication rate of each gene is about 1.06×10^{-4}/Myr. The gene duplication occurs randomly from the existing genes with a small degree. For another, creating orphan genes: the occurrence rate of an orphan gene is about 2.33×10^{-1}/Myr.

(ii) Deletion of interactions: the rate of interaction loss for each duplication gene is about 2.21×10^{-3}/Myr. On average, about 86.22% of the common interactions inherited from their parent genes are lost randomly for each duplication gene.

(iii) Addition of interactions: For each duplication gene, firstly, the rate of parent-child interaction for each duplication gene is about 4.62×10^{-4}/Myr. On average, only about 18.0% of duplicate genes shows the property of parent-child interaction. Secondly, the rate of new interaction for each duplication gene is about 6.54×10^{-2}/Myr. In general, the average new interaction of each duplication gene is about 26. The new duplicate genes develop new links through a 'rich-gets-richer' mechanism, the average degree of new interaction partners is about 161, and more than 80% of the new interaction partners have a degree larger than 35. For an orphan gene, the rate of new interaction for each orphan gene is about 4.57×10^{-2}/Myr. The attachment of an orphan gene also keeps a 'rich-gets-richer' mechanism, the average degree of an orphan gene is about 18, and the average degree of their new interaction partners is 133, more than 80% of the new interaction partners have a degree larger than 35.

Impact of younger genes on the topological structure of ancestral network

When the younger genes in branch 1–13 are integrated into the initial network G_0 branch by branch, we find that the number of interaction genes in the whole network would gradually increase (Figure 10A). The average degree of interaction genes would also gradually increase and eventually reach to a saturation state (Figure 10B).

Moreover, with the increase of evolutionary time, it is found that the average shortest path length of the whole network would gradually increase, but the increase is not too much (Figure 10C). The average clustering coefficient of the whole network also gradually increases (Figure 10D). This implies that with the gradual integration of younger genes into the ancestral network, the genes have a stronger and stronger tendency to form clusters or groups.

In particular, with the increase of evolutionary time, we can see that both the average node betweenness and average edge betweenness of the whole network would gradually increase (Figure 10E and F). This implies that with the gradual integration of younger genes into the network, the pivotability of every gene and each link becomes more and more strong. In addition, with the gradual integration of younger genes into the network, we find that both the total number of disconnected components in the whole network and the size of largest connected component would gradually increase (Figure 10G and H).

However, with the gradual integration of younger genes into the ancestral network, the exponent of degree distribution of the whole network does not change significantly (Figure 10I), which indicates that the degree distribution of

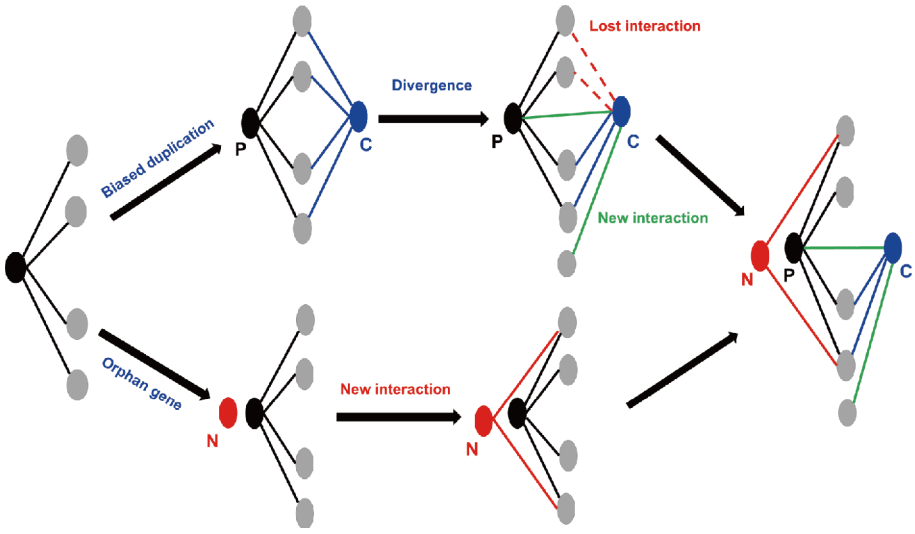

Figure 9 A flow diagram illustrating the evolutionary process of human gene coexpression network.

younger genes in different branches always follows a power-law distribution, that is, many younger genes with only a few links and a few younger genes with large numbers of links. The power-law degree distribution is a robust feature of the human gene coexpression network. Interestingly, the modularity coefficient of the whole network also does not change obviously (Figure 10J). This implies that the network always keeps a significant community structure. There always exists a good division of the network into small communities, and different communities show different functions.

By the above analysis, we can say that with the gradual integration of younger genes into the network, most of the topological structure features of the network would gradually increase. However, the exponent of degree distribution and modularity coefficient of the whole network do not change significantly, which implies that during the process of integration of younger genes into the ancestral network, the degree distribution always follows a power-law distribution and the network always keeps a community structure.

DISCUSSION

Uncovering the generic topological structure and organizing principles of human gene coexpression network is fundamental to our understanding of the real gene regulatory network (Barkai and Leibler, 1997; Albert et al., 2000; Barabási and Oltvai, 2004; Crombach and Hogeweg, 2008; Li et al., 2014). In this paper, firstly, we analyze the global structure features of a human gene coexpression network. Our results show that this network consists of 334 small components and one "giant" connected subnet comprising of 6,317 interacting genes. The degree distribution of this network approximates a power-law distribution, where the degree exponent is 1.264 (95% CI, 1.221–1.306). Interestingly, it is found that the degree distribution of younger genes (branch 1–13) also follows a power-law distribution. This indicates that many younger genes with only a few links and a few younger genes with large numbers of links. Some younger genes may become hub genes with critical phenotypic effects. The topology of a human gene coexpression network, derived from tissue-specific expression profiles, shows scale-free property, implying that the evolutionary self-organization of the network may follow a preferential principle. However, it needs to further explore whether a power-law degree distribution conveys any kind of advantage to human and whether natural selection probably favors the survival of organisms whose cellular networks have this degree distribution. Whether the power-law degree distribution can endow a network with robustness against perturbations still needs to be verified (Barkai and Leibler, 1997; Albert et al., 2000).

Moreover, we find that the average shortest path length of this network is 4.2, which implies that this network shows the property of small-world. The distribution of shortest path length of this network approximates a Poisson distribution. In particular, when the younger genes in branch 1–13 are integrated into the initial network branch by branch, we find that the average shortest path length of the whole network would gradually increase, but the increase is not too much. This indicates that there may exist certain evolutionary mechanisms that have maintained the average path length during evolution, which also needs further study.

In addition, it is found that the average clustering coefficient of whole network is 0.6019, and the average clustering coefficient of younger genes is larger than that of the elderly genes, which implies that the younger genes have a much stronger tendency to form clusters or groups. With the in-

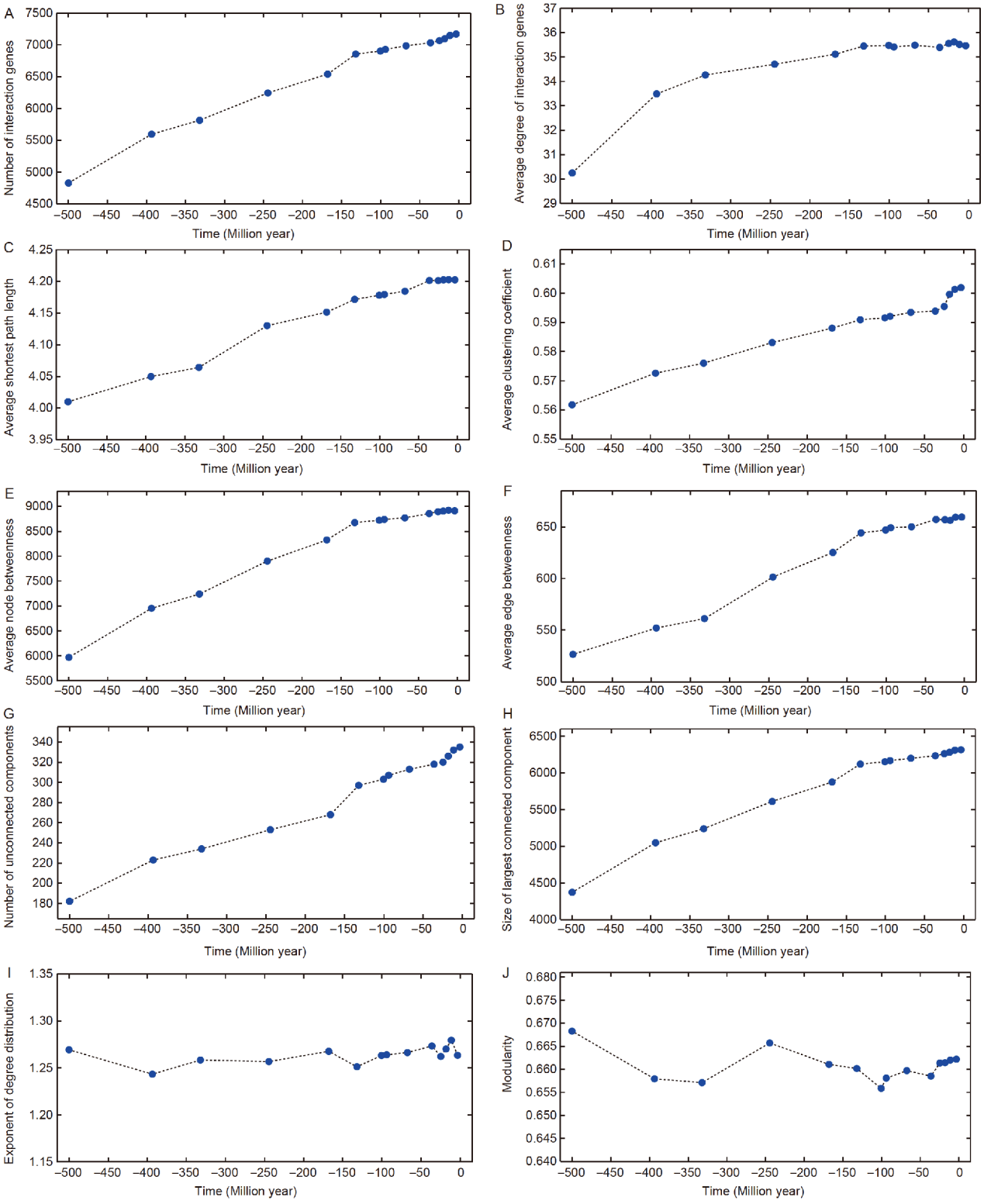

Figure 10 The impact of younger genes on the topological structure of ancestral gene coexpression network. A, The influence of younger genes on the number of interaction genes in the whole network. B, The influence of younger genes on the average degree of interaction genes. C, The influence of younger genes on the average shortest path length of the whole network. D, The influence of younger genes on the average clustering coefficient of the whole network. E, The influence of younger genes on the average node betweenness of the whole network. F, The influence of younger genes on the average edge betweenness of the whole network. G, The influence of younger genes on the total number of disconnected components in the whole network. H, The influence of younger genes on the size of largest connected component. I, The influence of younger genes on the exponent of degree distribution of the whole network. J, The influence of younger genes on the modularity coefficient of the whole network.

crease of gene age, their average clustering coefficient would gradually decrease. Particularly, we find that the genes with a larger degree show a property of hierarchical architecture and the younger genes with a larger degree also show a similar property of hierarchical architecture. This indicates that the communication between the different highly clustered neighborhoods may be maintained by a few hub genes (including younger genes). At the meantime, we find that the modularity coefficient of the whole network is larger than 0.65, which implies that the network may have a significant community structure, there may exist a good division of the network into small communities, different communities may show different functions, but it still needs to develop relevance for the experimental biologist, helping to elucidate the role of individual molecules in various cellular processes. More interestingly, we find that the average edge betweenness of younger genes is a little larger than that of the old genes, which further indicates that the younger genes and their new links play an important role in the overall pivotability of the network, the removal of a link induced by a younger gene may significantly affect the communication between many pairs of genes through the shortest paths between them. After gene duplication, compared with their parental genes, it is found that most duplicate genes eventually lose all common interactions, only 15.67% (205/1,308) of duplicate gene pairs share a common interaction partner. This implies that the human gene coexpression network contains few redundant duplicate genes.

Secondly, we explore the evolutionary model of the human gene coexpression network. Our results show that gene duplication and orphan genes are two dominant evolutionary force in shaping this coexpression network. Both the duplicate gene and orphan gene develop new links through a "rich-gets-richer" mechanism, that is, new genes prefer to connect to genes that already have many links, a process that is known as preferential attachment. Indeed, if a gene has many links, new genes will tend to connect to it with a higher probability. By using our proposed model, the global statistical properties of the network and its topologies may be well represented. However, in order to obtain a real gene regulatory network, it may be possible to infer general growth and design principles from the comprehensive properties of the network. Interestingly, with the gradual integration of younger genes into the ancestral network, we find that most of the topological structure features of the network would gradually increase (such as the average degree, the average clustering coefficient, the average node betweenness and average edge betweenness). However, the exponent of degree distribution and modularity coefficient of the whole network do not change significantly. This indicates that during the evolutionary process of human gene coexpression network, the degree distribution always follows a power-law distribution and the network always keeps a community structure. There may exist certain evolutionary mechanisms that have maintained the gene community structure during evolution, which needs further study. The observations made here stimulate a multitude of questions regarding their evolutionary significance. Is the change in network structure driven by neutral evolution or by natural selection for advantageous interaction patterns? Does the change in network configurations over time reflect different environments or different adaptations that the organism evolved at different times? It is also worth our further consideration that can one understand the network's global structure from a few key parameters, such as the rate of gene duplication and the rate of interaction loss (Wagner, 2001)?

This study has several limitations. First, in this gene coexpression network the strength of connections is ignored, that is, the network is unweighted. But for a real gene coexpression network, the strength of connections is different and the correlation between genes may be positive or negative. Second, when we consider the evolutionary process of gene coexpression network, we assume that the parental genes and existing genes in the network do not lose their interactions during evolution. More realistically, it is also likely that a certain amount of "rewiring" takes place during growth for the parental genes and existing genes in the network. Finally, except for the average degree, further development of this evolutionary model should consider different topological properties of genes and the underlying universal principle of gene-gene interactions.

We conclude from our results and their limitations that, the degree distribution of this network approximates a power-law distribution, this network shows the small-world property, has a hierarchical architecture and an obvious community structure. The younger genes may play an important role in the overall pivotability of the network. The network contains few redundant duplicate genes. Gene duplication and orphan genes are two dominant evolutionary forces in shaping this coexpression network. Both the duplicate gene and orphan gene develop new links through a 'rich-gets-richer' mechanism. With the gradual integration of younger genes into the network, most of the topological structure features of the network will gradually increase. However, the exponent of degree distribution and modularity coefficient of the whole network do not change significantly. The evolution of gene coexpression networks maintains the hierarchical and modular structures in human ancestors. Although more complex models will be needed, there is real biological content to the evolutionary model considered here that will likely provides some insights into the processes of evolution and the way in which biological networks function. This developing framework will significantly alter our understanding of gene networks and, eventually, will have important implications for the practice of medicine. It must be acknowledged that structure, topology, network usage, ro-

bustness and function are deeply interlinked, forcing us to complement the 'local' molecule-based research with integrated approaches that address the properties of the cell as a whole. Future progress is expected in many directions, ranging from the development of new theoretical methods to characterize the network topology to insights into the dynamics of motif clusters and biological function (Barabási and Oltvai, 2004; Oldham et al., 2006; Prieto et al., 2008; Sorrells and Johnson, 2015).

MATERIALS AND METHODS

Human gene coexpression data

The human gene correlation table is downloaded from http://coxpresdb.jp/download.shtml (COXPRESdb v. 6.0) (Obayashi et al., 2012). In this study, pairwise correlations between gene expression patterns were used to derive the gene coexpression network. The co-expressed gene table was calculated for 19,803 genes in human, by using 73,083 Gene Chip expression profiling data of 63 human tissues in NCBI GEO and using the method of Pearson correlation coefficients (PCC). If the absolute value of PCC of pairwise genes is greater than or equal to 0.5, the gene pairs are considered to be linked. In this way, we construct a human gene coexpression network (HGCN) that is relatively strong correlation. The detailed information about this gene coexpression network can be found in File S1 in Supporting Information (Obayashi et al., 2012).

Human gene age data

The updated dataset of human gene age is downloaded from http://gentree.ioz.ac.cn/download.php (Zhang et al., 2010, 2011). In this study, based on UCSC syntenic genomic alignment, each gene was dated and given branch assignment by inferring the absence and presence of orthologs along the vertebrate phylogenetic tree. There are 14 (0–13) branches in current age dating system. The younger the gene is, the bigger the branch number it is assigned. Here we use the dataset of "Highly Reliable Age Inference". The detailed information about this gene age dataset can be found in File S2 in Supporting Information (Zhang et al., 2010, 2011).

Gene origination mechanism data

The updated data of human gene origination mechanism is also downloaded from http://gentree.ioz.ac.cn/download.php (Zhang et al., 2010, 2011). The inference of origination mechanism is based on all-against-all BLASTP search for human proteins. Briefly, young genes are classified as DNA-level duplicates (D), RNA-level duplicates (retro genes, R) and orphan genes (A). We use the dataset of "Highly Reliable Mechanism Inference". The detailed information about this dataset can be found in File S3 in Supporting Information (Zhang et al., 2010, 2011).

Divergence time of species

The divergence time of two species is retrieved from http://www.timetree.org/ (Zhang et al., 2010, 2011; Hedges et al., 2015). In this database, the divergence time of two species or higher taxa is determined from thousands of published studies.

Graph analysis

If the absolute value of Pearson correlation coefficient (PCC) r of a pairwise genes is greater than or equal to 0.5, that is, $|r|\geqslant 0.5$, then the gene pairs are considered to be coexpressed and linked. In this way, we construct a human gene coexpression network (HGCN). The gene coexpression network corresponds to a graph, and each gene corresponds to a node, each link corresponds to an edge. Here we analyze the most basic network measures that allow us to understand the global structure features of human gene coexpression network and their evolutionary process.

Average degree

The most basic characteristic of a gene (node) in the network is its degree (or connectivity), k, which tells us how many links the gene has to other genes. An undirected network with N genes and M links is characterized by an average degree $<k>=2M/N$, (where $<k>$ denotes the average degree).

Degree distribution

The degree distribution, $P(k)$, gives the probability that a selected gene has exactly k links. $P(k)$ can be obtained by counting the number of genes $N(k)$ with degree k and dividing by the total number of genes N in the network. The degree distribution allows us to distinguish different types of networks. A power-law degree distribution $P(k)=\alpha k^{-\beta}$ indicates that a few hub genes hold together numerous small degree genes. The value of degree exponent β determines many properties of the network. The smaller the value of β, the more important the role of hub genes in the network.

Average shortest path length

Distance in the network is measured with the path length, which tells us how many links we need to pass through to travel between two selected genes. Generally, there are many alternative paths between two selected genes, the shortest path means the path with the minimal number of links be-

tween the two selected genes. Sometimes there is no path between two genes. The average shortest path length, $<l>$, represents the average over the shortest paths between all pairs of genes and offers a measure of a network's overall navigability.

The distribution of shortest path length, $D(l)$, gives the probability that a selected pair of genes has exactly shortest path length l. $D(l)$ can be calculated by counting the number of shortest paths $P(l)$ with length l and dividing by the total number of shortest paths P.

Average clustering coefficient

In order to define the clustering coefficient $C(v)$ of a gene v, consider all k_v genes adjacent to a gene v, and count the number of links m_v that exist among these k_v genes (not including links connecting them to v). It can be seen that the maximum possibility of m_v is $k_v(k_v-1)/2$, in which all m_v genes are connected to each other. Therefore, the clustering coefficient $C(v)$ of a gene v is defined as $C(v)=m_v/(k_v(k_v-1)/2)$, which measures the "cliquishness" of the neighborhood of gene v, that is, what fraction of the genes adjacent to gene v are also adjacent to each other (Watts and Strogatz, 1998). In some network there may be no clustering coefficient for a gene. In an extension, the average clustering coefficient of a network, $<C>$, is defined as the average of $C(v)$ over all genes v, which characterizes the overall tendency of genes to form clusters or groups. Furthermore, an important measure of the network's structure is the function $C(k)$, which is defined as the average clustering coefficient of all genes with k links. If $C(k)\sim k^{-1}$, then the network may have a hierarchical architecture. The function $C(k)$ is independent of the network's size and therefore may capture a network's generic features.

Modularity

The modularity of a network with respect to some division (or gene communities) measures how good the division is, or how separated are the different gene communities from each other, in the sense that there are many links within communities and only a few between them. It is defined as

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j),$$

where m is the number of edges, A_{ij} is the element of the adjacency matrix A in row i and column j, k_i is the degree of gene i, k_j is the degree of gene j, c_i is the type (or component) of gene i, c_j is that of gene j, the sum goes over all i and j pairs of genes, and $\delta(c_i,c_j)$ is 1 if $c_i=c_j$ and 0 otherwise. If the fraction of within-community edges is no different from what we would expect for the randomized network, then this modularity will be zero. Nonzero values represent deviations from randomness, and in practice it is found that a value above about 0.3 is a good indicator of significant community structure in a network. If high values of the modularity correspond to good divisions of a network into communities, then one should be able to find such good divisions by searching through the possible candidates for ones with high modularity (Clauset et al., 2004).

Average node betweenness

The betweenness of a gene v is defined by the number of shortest paths going through the gene v, which is calculated by

$$B(v) = \sum_{i\neq j\neq v}\frac{g_{ivj}}{g_{ij}},$$

where g_{ij} is the total number of shortest paths from gene i to gene j and g_{ivj} is the number of those shortest paths that pass through the gene v. The node betweenness provides a good measure of a gene's overall pivotability. The average node betweenness of a network, $<B_1>$, is defined as the average of $B(v)$ over all genes v.

Average edge betweenness

The edge betweenness is defined as the number of the shortest paths that go through an edge (link) E in a graph or network (Girvan and Newman, 2002), which is calculated by

$$B(e) = \sum_{i\neq j}\frac{g_{iEj}}{g_{ij}},$$

where g_{ij} is the total number of shortest paths from gene i to gene j and g_{iEj} is the number of those shortest paths that pass through the link E. A link with a high edge betweenness represents a bridge-like connector between two parts of a network, and the removal of which may affect the communication between many pairs of genes through the shortest paths between them. The average edge betweenness of a network, $<B_2>$, is defined as the average of $B(E)$ over all edges E. All of these calculations and visualization can be implemented on R platform, by using a network analysis R package referred to as "igraph".

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

Acknowledgements *We thank Profs. Yicang Zhou and Yanni Xiao for their valuable discussion. This work was supported by grants from the National Natural Science Foundation of China (11571272, 11201368 and 11631012), the National Science and Technology Major Project of China (2012ZX10002001), the Natural Science Foundation of Shaanxi Province (2015JQ1011) and the China Postdoctoral Science Foundation (2014M560755).*

References

Albert, R., Jeong, H., and Barabási, A.L. (2000). Error and attack tolerance of complex networks. Nature 406, 378–382.

Barkai, N., and Leibler, S. (1997). Robustness in simple biochemical networks. Nature 387, 913–917.

Barabási, A.L., and Albert, R. (1999). Emergence of scaling in random networks. Science 286, 509–512.

Barabási, A.L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nat Rev Genet 5, 101–113.

Chung, F., Lu, L., Dewey, T.G., and Galas, D.J. (2003). Duplication models for biological networks. J Comput Biol 10, 677–687.

Cohen, R., and Havlin, S. (2003). Scale-free networks are ultrasmall. Phys Rev Lett 90, 058701.

Clauset, A., Newman, M.E.J., and Moore, C. (2004). Finding community structure in very large networks. Phys Rev E 70, 066111.

Chung, W.Y., Albert, R., Albert, I., Nekrutenko, A., and Makova, K.D. (2006). Rapid and asymmetric divergence of duplicate genes in the human gene coexpression network. BMC Bioinformatics 7, 46, 1–14.

Crombach, A., and Hogeweg, P. (2008). Evolution of evolvability in gene regulatory networks. PLoS Comput Biol 4, e1000112.

Chen, S., Krinsky, B.H., and Long, M. (2013). New genes as drivers of phenotypic evolution. Nat Rev Genet 14, 645–660.

Girvan, M., and Newman, M.E.J. (2002). Community structure in social and biological networks. Proc Natl Acad Sci USA 99, 7821–7826.

Horvath, S., and Dong, J. (2008). Geometric interpretation of gene coexpression network analysis. PLoS Comput Biol 4, e1000117.

Hedges, S.B., Marin, J., Suleski, M., Paymer, M., and Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. Mol Biol Evol 32, 835–845.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabási, A.L. (2000). The large-scale organization of metabolic networks. Nature 407, 651–654.

Jordan, I.K., Marino-Ramirez, L., Wolf, Y.I., and Koonin, E.V. (2004). Conservation and coevolution in the scale-free human gene coexpression network. Mol Biol Evol 21, 2058–2070.

Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J., and Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. Genome Res 14, 1085–1094.

Li, M., Li, Q., Ganegoda, G.U., Wang, J.X., Wu, F.X., and Pan, Y. (2014). Prioritization of orphan disease-causing genes using topological feature and GO similarity between proteins in interaction networks. Sci China Life Sci 57, 1064–1071.

Maslov, S., and Sneppen, K. (2002). Specificity and stability in topology of protein networks. Science 296, 910–913.

Oldham, M.C., Horvath, S., and Geschwind, D.H. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proc Natl Acad Sci USA 103, 17973–17978.

Obayashi, T., Okamura, Y., Ito, S., Tadaka, S., Motoike, I.N., and Kinoshita, K. (2012). COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals. Nucleic Acids Res 41, D1014–D1020.

Pastor-Satorras, R., Smith, E., and Solé, R.V. (2003). Evolving protein interaction networks through gene duplication. J Theor Biol 222, 199–210.

Prieto, C., Risueño, A., Fontanillo, C., and De las Rivas, J. (2008). Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. PLoS ONE 3, e3911.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., and Barabási, A.L. (2002). Hierarchical organization of modularity in metabolic networks. Science 297, 1551–1555.

Shen-Orr, S.S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat Genet 31, 64–68.

Sorrells, T.R., and Johnson, A.D. (2015). Making sense of transcription networks. Cell 161, 714–723.

Tautz, D., and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. Nat Rev Genet 12, 692–702.

Watts, D.J., and Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. Nature 393, 440–442.

Wagner, A. (2001). The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol Biol Evol 18, 1283–1292.

Wagner, A. (2003). How the global structure of protein interaction networks evolves. Proc R Soc London Ser B-Biol Sci 270, 457–466.

Yu, H., Mitra, R., Yang, J., Li, Y.Y., and Zhao, Z.M. (2014). Algorithms for network-based identification of differential regulators from transcriptome data: a systematic evaluation. Sci China Life Sci 57, 1090–1102.

Zhang, Y.F., Zhang, R., and Su, B. (2009). Diversity and evolution of microRNA gene clusters. Sci China Ser C-Life Sci 52, 261–266.

Zhang, Y.E., Vibranovski, M.D., Landback, P., Marais, G.A.B., and Long, M. (2010). Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. PLoS Biol 8, e1000494.

Zhang, Y.E., Landback, P., Vibranovski, M.D., and Long, M. (2011). Accelerated recruitment of new brain development genes into the human genome. PLoS Biol 9, e1001179.

Zhang, W., Landback, P., Gschwend, A.R., Shen, B., and Long, M. (2015). New genes drive the evolution of gene interaction networks in the human and mouse genomes. Genome Biol 16, 202.

SUPPORTING INFORMATION

File S1 Pearson correlation coefficients (PCC), gene age (Branch 0–13) and human gene coexpression network (Obayashi et al., 2012).

File S2 Human gene age dataset, "Highly Reliable Age Inference" (Zhang et al., 2010, 2011).

File S3 Human gene origination mechanism dataset, "Highly Reliable Mechanism Inference" (Zhang et al., 2010, 2011).

The supporting information is available online at http://life.scichina.com and https://link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.