Dissecting the Human Protein-Protein Interaction Network via Phylogenetic Decomposition

Cho-Yi Chen[1], Andy Ho[2], Hsin-Yuan Huang[3], Hsueh-Fen Juan[1,2,*], Hsuan-Cheng Huang[4,*]

[1]Genome and Systems Biology Degree Program, Department of Life Science, Institute of Molecular and Cellular Biology, [2]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan, [3]Taipei Municipal Jianguo High School, Taipei, Taiwan, [4]Institute of Biomedical Informatics, Center for Systems and Synthetic Biology, National Yang-Ming University, Taipei, Taiwan

*E-mail: yukijuan@ntu.edu.tw; hsuancheng@ym.edu.tw

Supplementary Information

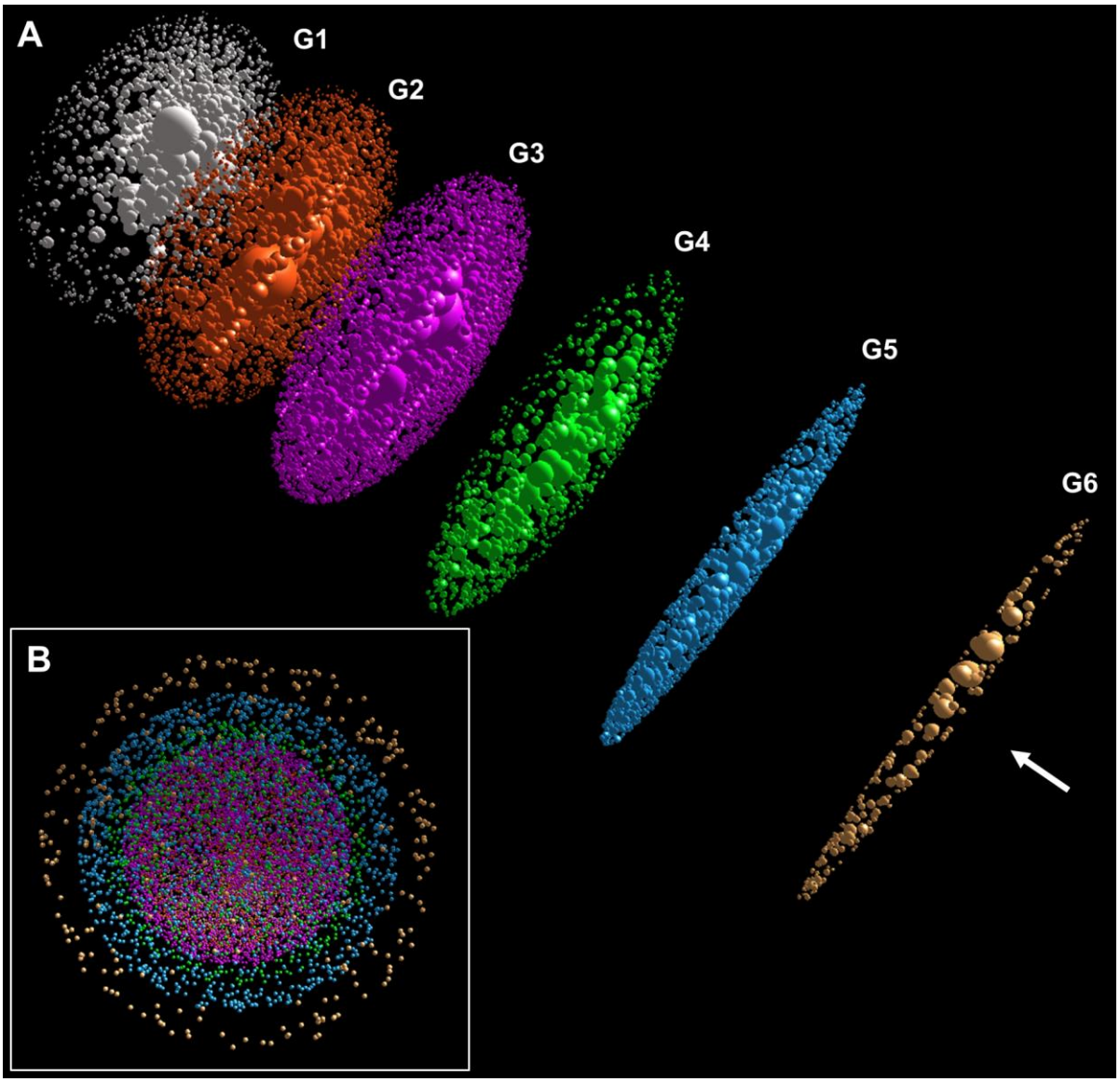


Figure S1. 3D view of phylo-decomposed human PPI network. (A) A 45-degree-angle view. Node size reflects node degree. The arrow indicates the view angle for (B). Node size was reset to equal unit size in (B) for visual clarity.

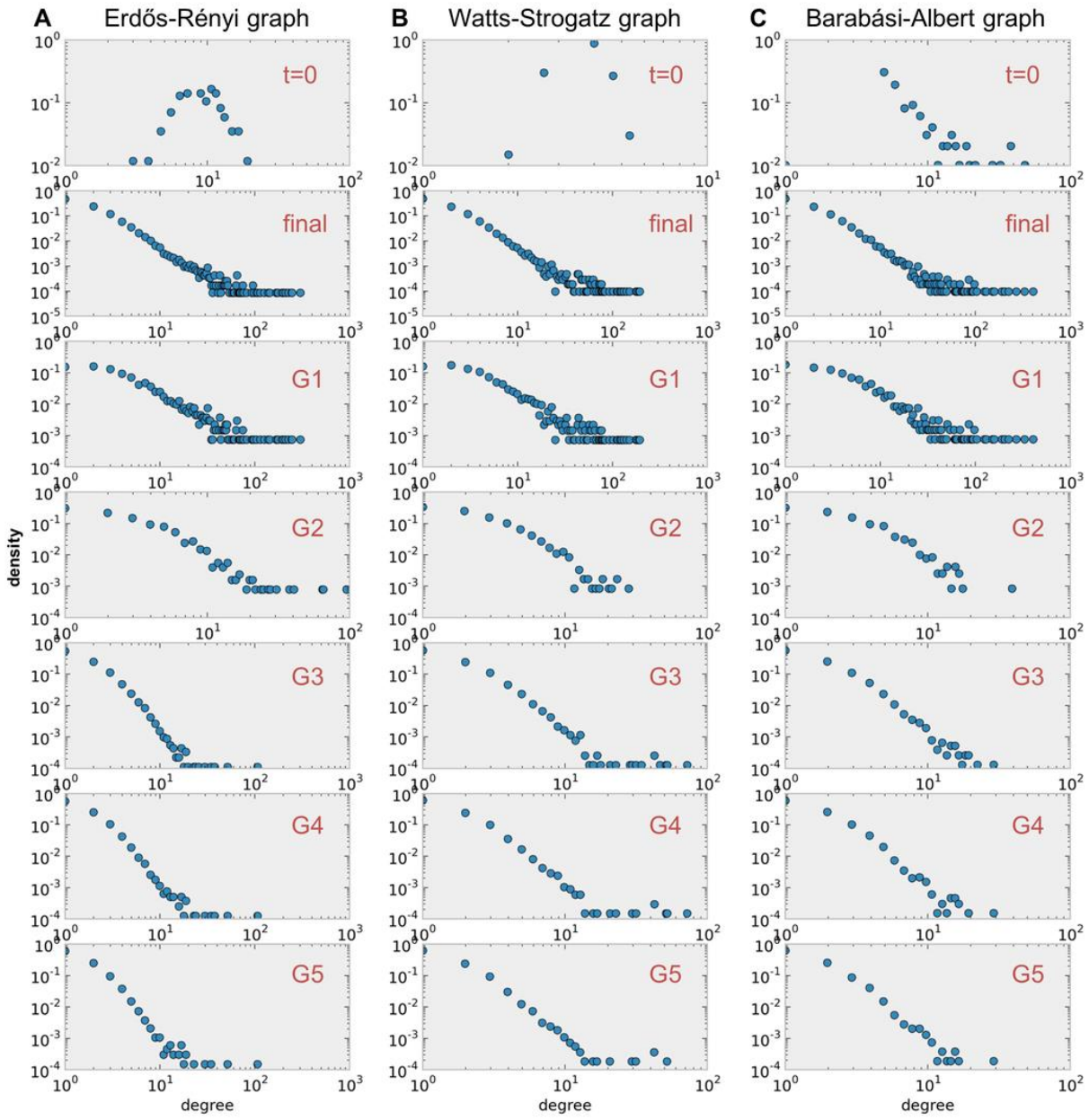


Figure S2. Network growth simulation based on the perturbation-avoidance model. Three different kinds of initial graphs were used as seed networks: (A) Erdős-Rényi graph, (B) Watts-Strogatz small-world graph, and (C) Barabási-Albert scaling graph. We stopped the growth when the network size reached 10,000 nodes. The approximate power-law distribution of node connectivity (degree) can be observed on the log-log plots (t=0: the initial networks, final: the final networks reaching 10,000 nodes). This behavior is ubiquitous across all age groups (G1-G5, each group contains ~2,000 nodes), regardless of initial network choices.

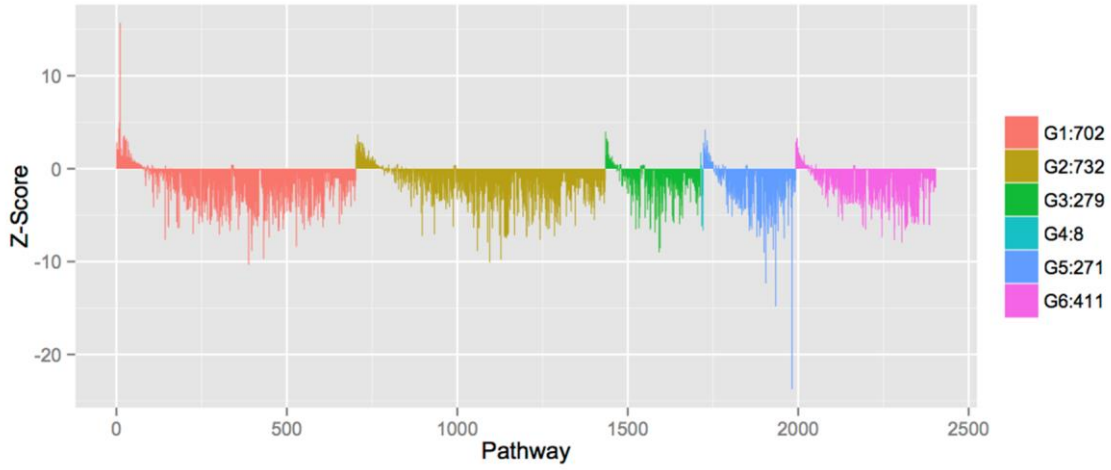


Figure S3. Biological pathways tend to be age homogeneous. We classified the KEGG pathways into different aged groups based on how many organisms possess the given pathways, following similar assignment criteria that we had used to classify human proteins. After that, we computed the age SD for proteins engaging in each pathway (as a measure of age homogeneity). Subsequently, the Z-scores of observed degrees of age homogeneity were estimated via Monte-Carlo resampling procedures. Low Z-scores indicate more significant age homogeneity. Pathways in each group are ranked in descending order of age SD. Generally, all pathways exhibit strong age homogeneity.

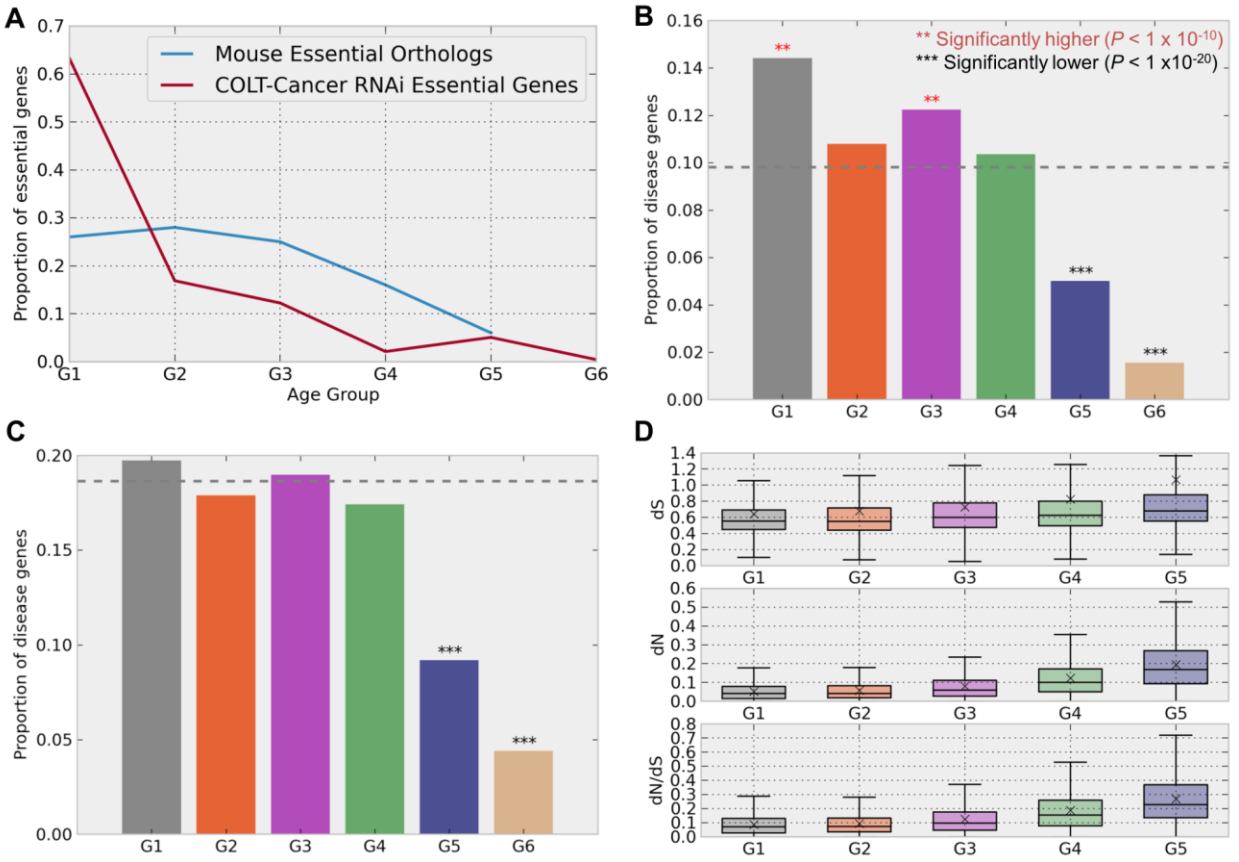


Figure S4. Gene essentiality, disease susceptibility, and evolutionary rates of age groups. (A) Distribution of essential genes across age groups. Two different sources of gene essentiality annotation were used. (B-C) Distribution of disease-susceptible genes across age groups. The lists of genes associated with genetic disorders were obtained either from HPRD (B) or OMIM (C). Dashed lines indicate the global means. *P* values were derived from Fisher's exact test. (D) Human-mouse evolutionary rates of age groups. Cross marks indicate the means. dN: the number of non-synonymous substitutions per non-synonymous site; dS: the number of synonymous substitutions per synonymous site.

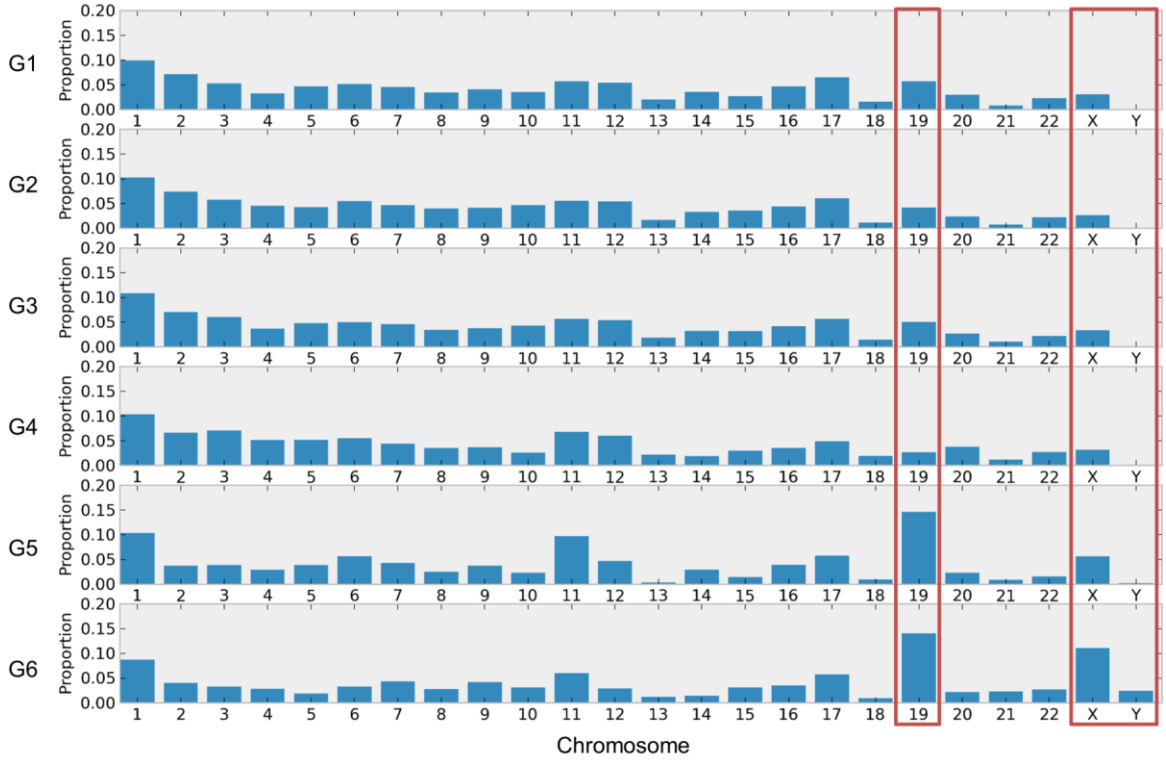


Figure S5. Human chromosome distribution of age groups. For better comparison of distribution patterns, all distributions were normalized by the size of the corresponding categories. Chromosomes 19, X, and Y were highlighted for further discussion in the main text.

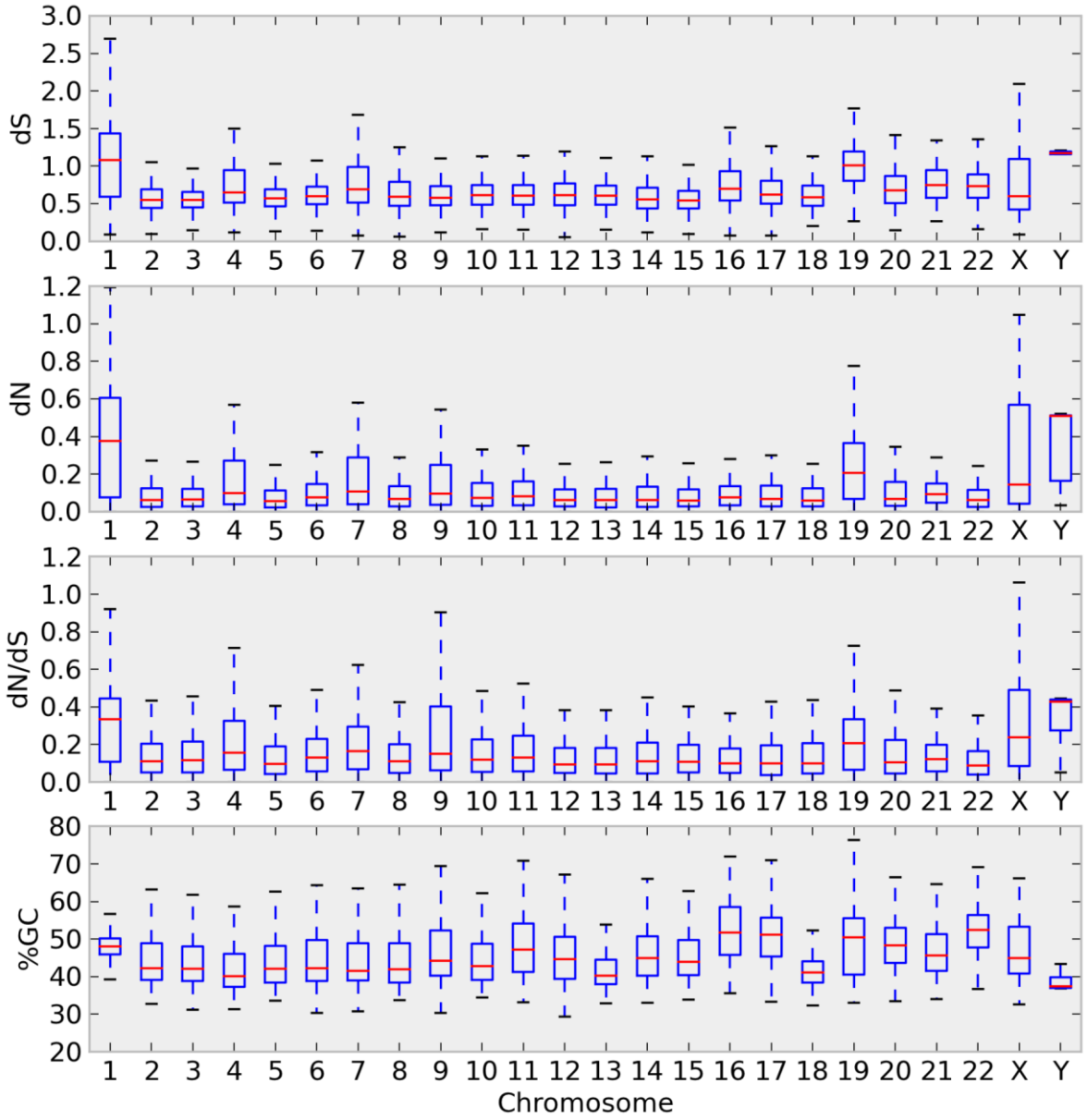
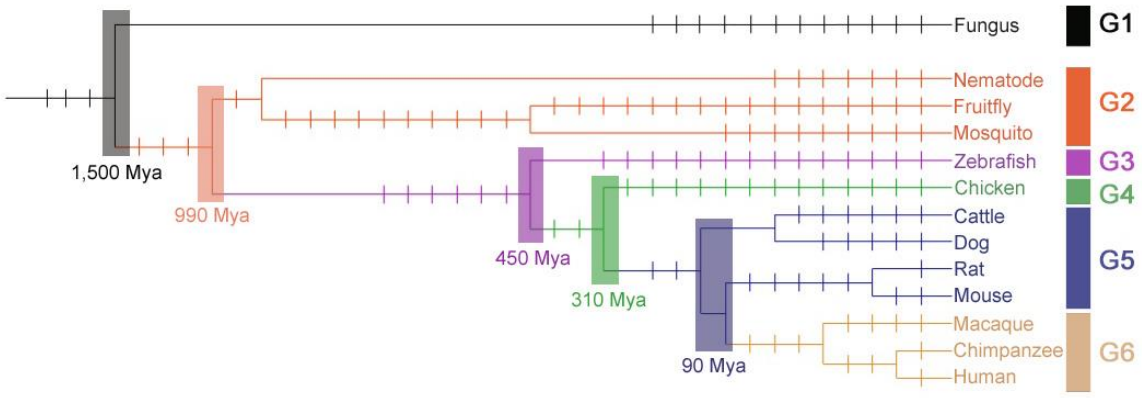


Figure S6. Human-mouse gene evolutionary rates and GC contents across human chromosomes. All measurements were calculated from protein-coding genes. dN: the number of non-synonymous substitutions per non-synonymous site; dS: the number of synonymous substitutions per synonymous site; %GC: GC contents of genes. All information was obtained via Ensembl Biomart.



Network Context

	Global		Maximum Connected Component	
Context	**Nodes**	**Edges**	**Nodes**	**Edges**
P = HPRD∪DIP∪IntAct∪BioGRID∪MINT*	12104	82551		
P∩HomoloGene	11839	79549	11603	79362
G1 (Eukaryotes)	1784	5396	1749	5374
G2 (Metazoans)	2892	7401	2859	7387
G3 (Vertebrates)	4507	11998	4431	11946
G4 (Tetrapods)	972	558	945	543
G5 (Mammals)	1387	768	1330	734
G6 (Primates)	297	62	289	57

*All human PPIs in databases HPRD, DIP, IntAct, BioGRID, MINT prior to Aug. 2011, downloaded from PrePPI database

PPI Summary

Group	G1	G2	G3	G4	G5	G6
G1	5396	9326	10511	1613	1473	255
G2		7401	16356	2641	2291	318
G3			11998	3762	3413	412
G4				558	770	99
G5					768	126
G6						62

Table S1. Statistics of the human PPI network used in this study.

Top 3 most representative pathways in KEGG

G1	G2	G3	G4	G5	G6
Ribosome	Inositol phosphate metabolism	Neuroactive ligand-receptor interaction	Cell adhesion molecules (CAMs)	Olfactory transduction	Antigen processing and presentation
Spliceosome	Phosphatidylinositol signaling system	Pathways in cancer	Hematopoietic cell lineage	Systemic lupus erythematosus	Natural killer cell mediated cytotoxicity
Aminoacyl-tRNA biosynthesis	Insulin signaling pathway	MAPK signaling pathway	Cytokine-cytokine receptor interaction	Drug metabolism	Regulation of autophagy

Top 3 most representative domains in INTERPRO

G1	G2	G3	G4	G5	G6
NAD(P)-binding domain	Serine/threonine protein kinase-related	BTB/POZ-like	Fibronectin, type III-like fold	Olfactory receptor	Krueppel-associated box
DEAD-like helicase, N-terminal	Protein kinase, core	Homeobox, conserved site	EGF-like region, conserved site	GPCR, rhodopsin-like superfamily	Zinc finger, C2H2-type/integrase, DNA-binding
Helicase, superfamily 1 and 2, ATP-binding	Serine/threonine protein kinase	Leucine-rich repeat, typical subtype	EGF-like, type 3	7TM GPCR, rhodopsin-like	Keratin, high sulphur B2 protein

Table S2. Representative engaged pathways and protein domains.