Electronic Supplementary Information

Human genes with a greater number of transcript variants tend to show biological features of housekeeping and essential genes

Jae Yong Ryu,[a] Hyun Uk Kim[abd] and Sang Yup Lee[*abcd]

[a] Metabolic and Biomolecular Engineering National Research Laboratory, Department of Chemical and Biomolecular Engineering (BK21 Plus Program), Center for Systems and Synthetic Biotechnology, Institute for the BioCentury, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 305-701, Republic of Korea.

[b] BioInformatics Research Center, KAIST, Daejeon 305-701, Republic of Korea.

[c] BioProcess Engineering Research Center, KAIST, Daejeon 305-701, Republic of Korea.

[d] The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Hørsholm, Denmark.

[*] e-mail: leesy@kaist.ac.kr

Fig. S1 Features of exons in human genes. (A) A scatter plot depicting the correlation between the number of exons and the number of transcript variants in all the human genes (Pearson correlation coefficient = 0.39). (B) Distribution of the number of exons for the HK (*n* = 3,804), TS (*n* = 2,293), ES (*n* = 2,472), and NE genes (*n* = 3,811). Boxes represent the 25th-75th percentiles, while whiskers represent the 5th-95th percentiles. The line inside the box indicates the median value of the distribution. Abbreviations are: HK, housekeeping; TS, tissue-selective; ES, essential; NE, non-essential.

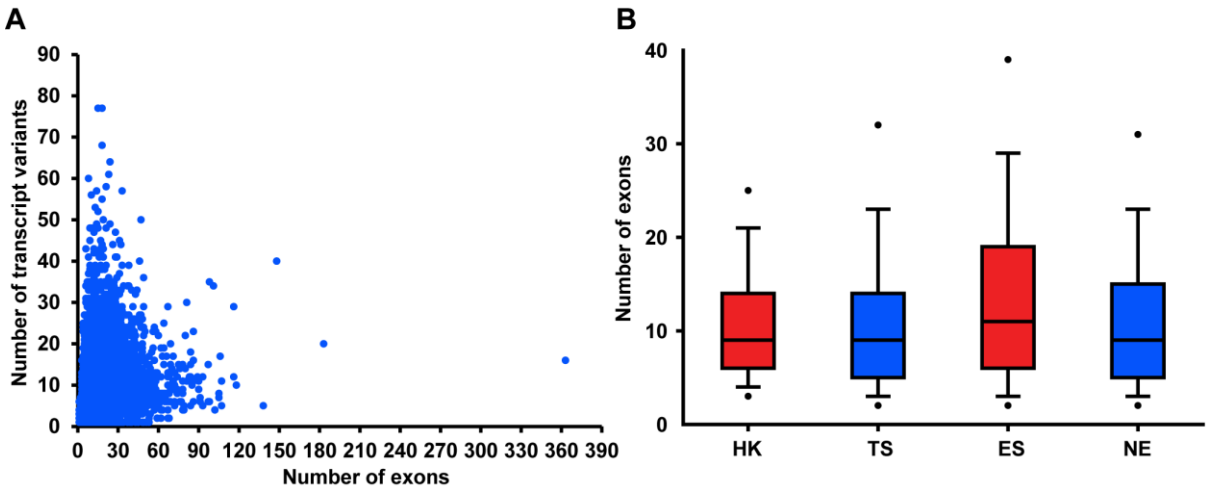


Fig. S2 A heat map showing percentage of the number of expressed genes in each group against each tissue. The HK, TS, ES, and NE genes were excluded for this analysis. Tissue-specific expression data were obtained from proteomics studies on 32 different human tissues.[1] The percentage represents the number of expressed genes among all the genes in each tissue. Tissue names are shown in the *x*-axis, and group names corresponding to the number of transcript variants are indicated on the *y*-axis. Some groups having horizontal blue lines (between the groups 36 and 64) represent that they do not have any genes left after exclusion of the HK, TS, ES, and NE genes. Abbreviations are: HK, housekeeping; TS, tissue-selective; ES, essential; NE, non-essential.

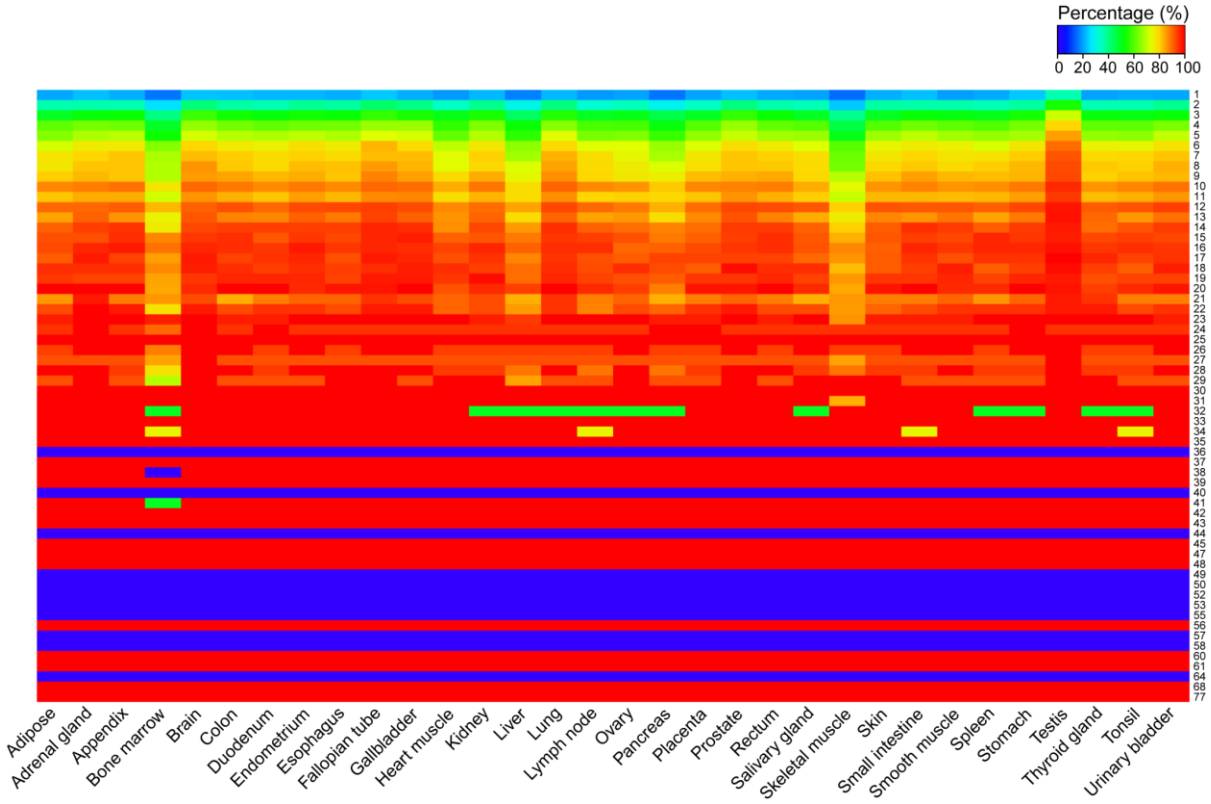


Fig. S3 A heat map showing the percentages of orthologs in gene groups having different number of transcript variants in 64 vertebrate species. Data on orthologs were obtained from OrthoDB.[2] The percentage represents the number of orthologs among all the genes present in the corresponding gene group and species. The HK, TS, ES, and NE genes were excluded for this analysis. The *x*-axis shows the vertebrate species, which were clustered according to their order; the order names are shown only if it has more than 3 relevant species. The *y*-axis is the group name corresponding to the number of transcript variants. Some groups having horizontal blue lines (between the groups 36 and 64) represent that they do not have any genes left after exclusion of the HK, TS, ES, and NE genes. Abbreviations are: HK, housekeeping; TS, tissue-selective; ES, essential; NE, non-essential.

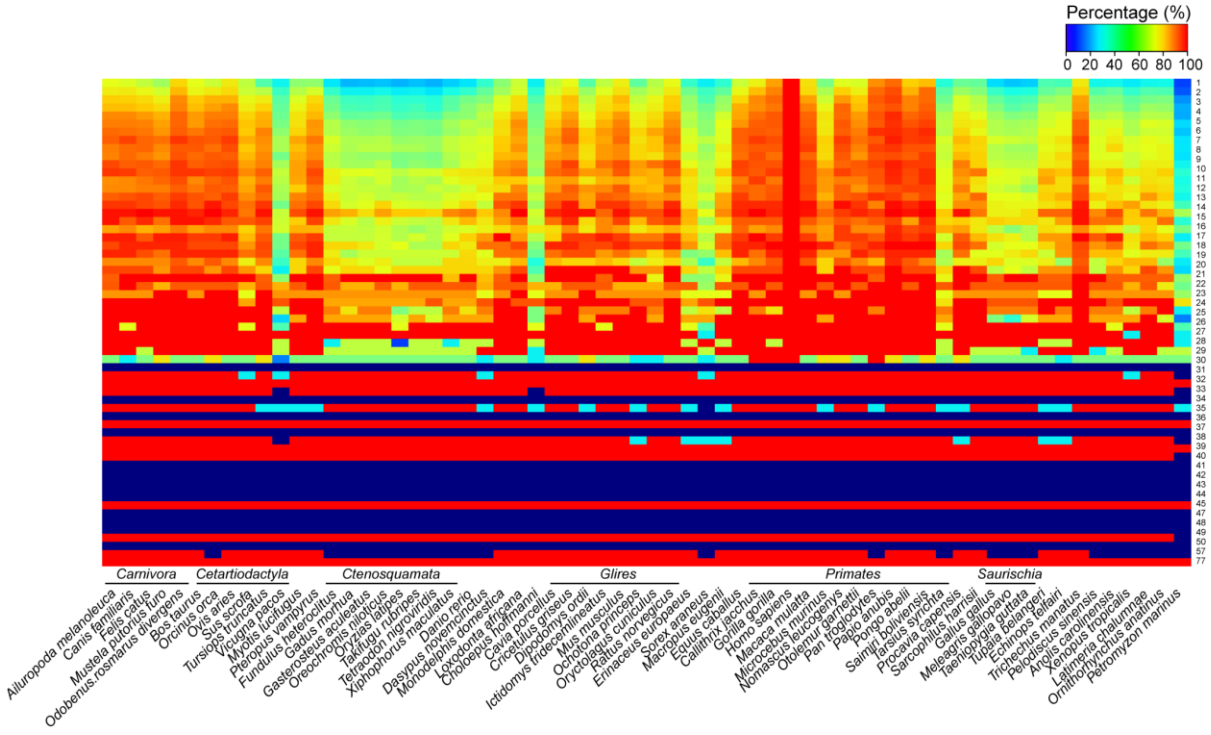
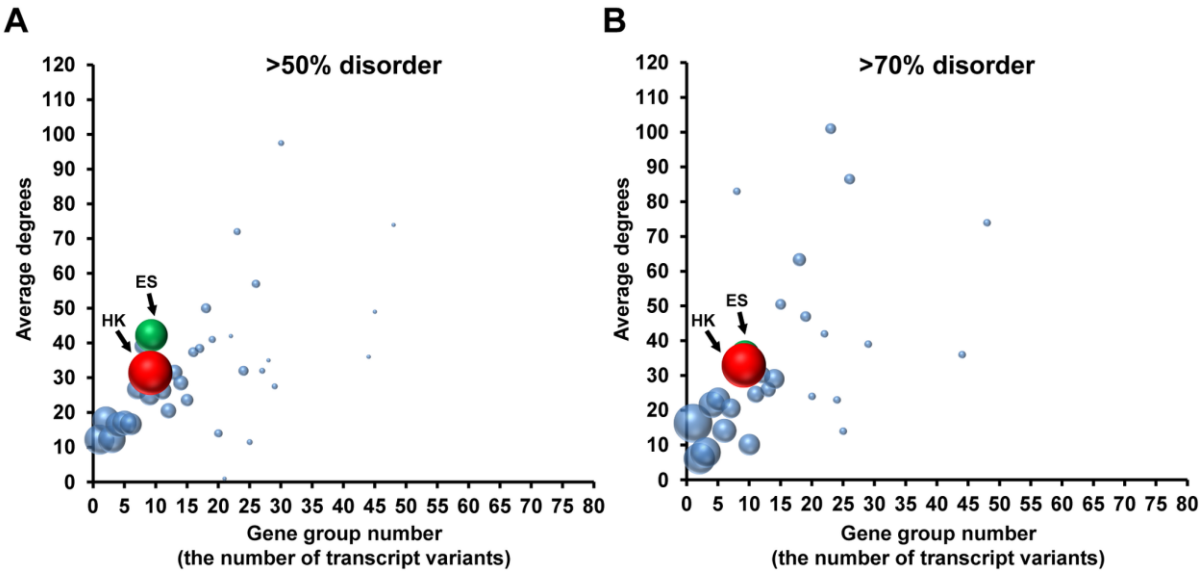


Fig. S4 Bubble plot showing the average degrees of proteins encoded by human genes in 60 groups, and the HK and ES genes in human PPI network. Protein interactome data were downloaded from PINA 2.0 database.[3] This network contains 17,109 nodes and 166,776 edges. The two bubble plots were presented for proteins with (A) >50% disorder only and (B) >70% disorder only. Red and green bubbles represent groups having the HK and ES genes, respectively. Blue bubbles represent 60 groups classified by the number of transcript variants. The bubbles for these genes are not shown because they block those of genes in 60 groups. Bubble size indicates the number of genes in each group. Abbreviations are: HK, housekeeping; ES, essential.



References

1. M. Uhlen, L. Fagerberg, B. M. Hallstrom, C. Lindskog, P. Oksvold, A. Mardinoglu, A. Sivertsson, C. Kampf, E. Sjostedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P. H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen and F. Ponten, *Science*, 2015, **347**, 1260419.
2. E. V. Kriventseva, N. Rahman, O. Espinosa and E. M. Zdobnov, *Nucleic Acids Res.*, 2008, **36**, D271-275.
3. M. J. Cowley, M. Pinese, K. S. Kassahn, N. Waddell, J. V. Pearson, S. M. Grimmond, A. V. Biankin, S. Hautaniemi and J. Wu, *Nucleic Acids Res.*, 2012, **40**, D862-865.