Estimating the selection pressure and evolutionary rate of proteins on the non-neutral hypothesis of synonymous mutations

Jiachen Ye[1], Chunmei Cui[1], Rui Fan[1], and Qinghua Cui[1,2*]

[1]Department of Biomedical Informatics, State Key Laboratory of Vascular Homeostasis and Remodeling, School of Basic Medical Sciences, Peking University, 38 Xueyuan Rd, Beijing, 100191, China

[2]School of Sports Medicine, Wuhan Institute of Physical Education, No. 461 Luoyu Rd. Wuchang District, Wuhan 430079, Hubei Province, China

* To whom the correspondence should be addressed:

Qinghua Cui, email: cuiqinghua@bjmu.edu.cn

Abstract

The Nonsynonymous/Synonymous substitution rate ratio (Ka/Ks) is a commonly used metric to estimate the selection pressure and evolutionary rate of proteins in comparative genomics, which plays critical roles in molecular evolution in both biology and medicine. A fundamental assumption of Ka/Ks is that synonymous mutations are evolutionarily neutral and not subject to natural selection as they do not alter protein sequences and function. However, a number of studies have demonstrated that synonymous mutations are non-neutral and may lead to diseases through a number of mechanisms, such as altering miRNA regulation. This further implies that synonymous mutations also participate in the process of natural selection and thus Ka/Ks should be redefined as well. For this purpose, here we propose an improved Ka/Ks ratio, *i*Ka/Ks, which re-computed the neutral substitution rate by taking the altered status of miRNA binding into consideration, and thereby incorporate the impact of synonymous mutations on miRNA regulation. As a result, *i*Ka/Ks shows better performance than Ka/Ks when comparing them using their correlation with expression distance. Moreover, case studies showed that *i*Ka/Ks is able to identify the positive/negative selection genes that are missed by Ka/Ks. For example, TMEM72/Tmem72 is estimated to be positively selected by *i*Ka/Ks (1.13) but negatively selected by the conventional Ka/Ks ratio (0.21). Further evidence showed its rapid evolution, which further support the power of the new algorithm.

Keywords: Ka/Ks, selection pressure, synonymous mutations, microRNAs

Introduction

The comprehension of species interrelations lies at the core of evolutionary biology[1]. Comparative genomics, as a cornerstone methodology in contemporary biomedical research[2], involves scrutinizing genetic information within and across organisms to unveil the evolutionary trajectory, structure, and functionality of genes, proteins, and non-coding RNAs[3]. Through systematically comparative analysis, the biological relationships and evolutionary dynamics among species or among molecules can be methodically explored and evaluated. This facilitates a deeper understanding of the structure and function of orthologous genes, thus enriching our comprehension of human diseases and their potential therapeutic targets[4].

Comparative genomics has provided a new means for understanding evolution by natural selection[5]. The ratio of nonsynonymous to synonymous substitution rate (Ka/Ks, also known as dN/dS) holds significantly importance for reconstructing phylogenies and elucidating the evolutionary dynamics of protein-coding sequences across closely related but diverged species[6,7]. The ratio serves as a pivotal metric for estimating the selection pressure and evolutionary rate of proteins and assessing the divergence among homologous gene sequences[8]. Generally, a diminished Ka/Ks value signifies heightened conservation within the homologous gene pair, with Ka/Ks << 1 indicating robust purifying selection. Conversely, an elevated Ka/Ks ratio signals increased disparity between homologous genes, with Ka/Ks > 1 suggestive of positive selection acting upon the gene. It is pertinent to note that the Ka/Ks algorithm is based on a fundamental assumption that synonymous mutations are neutral as they do not alter protein sequences and functions and thus are devoid of any influence on phenotypes and diseases[9]. According to this assumption, the synonymous substitution rate Ks

functions as a baseline for the background nucleotide substitution rate, and the Ka/Ks ratio delineates the strength of natural selection driving gene evolution[10].

However, increasing evidence has showed that synonymous mutations could be non-neutral[10,11]. They could even be considered to have comparable effects with nonsynonymous mutations in yeast[12]. Another study has brought to light the presence of codon usage bias at certain synonymous mutation sites, hinting at the potential non-neutrality of synonymous mutations and their consequential functional implications[13]. Notably, the association between miRNAs and synonymous mutations has emerged as a pivotal focus. This is because coding regions have been found to directly interact with miRNAs[14,15] and a proportion (40%-60%) of miRNA binding sites have been identified to locate within mRNA coding regions[16,17]. Synonymous mutations may lead to melanoma[18] and Crohn's disease[19] by altering miRNA binding. Wang *et al.* emphasized that nearly half of synonymous SNPs could modulate protein expression levels through miRNA-mediated gene regulation, leading to functional consequences and disease risks[20].

These observations indicate that synonymous mutations have the potential to modify miRNA binding sites, consequently impacting gene function. Thus, assuming synonymous substitution rate as the background nucleotide substitution rate is questionable, and thus the Ka/Ks algorithm should be re-defined. For this purpose, here we present an improved Ka/Ks algorithm, *i*Ka/Ks, by considering the alteration of miRNA binding by synonymous mutations. The results showed that the new algorithm demonstrates better performance compared with the conventional one.

Materials and methods

Data collection

We obtained the coding sequences of all orthologous genes between human (GRCh38) and mouse (GRCm39) from Ensembl (https://www.ensembl.org/biomart/martview/)[21], prioritizing the longest transcript sequence for genes with multiple transcripts. Subsequently, sequence alignment using MAFFT[22] yielded a total of 22,106 pairs of orthologous genes. Mature miRNA sequences were retrieved from miRBase v22[23].

Evaluation of the miRNA-synonymy of mutation sites

Each mutant site is categorized as either synonymous or nonsynonymous based on whether it alters the corresponding amino acid sequence. Drawing a parallel to this classification, we distinguish between miRNA-synonymous and miRNA-nonsynonymous sites according to their influence on miRNA binding on target mRNAs. For each mutation, a sliding window of 31-nt is employed to capture the sequence of the site along with its flanking 15-nt on both sides. Subsequently, we predict the miRNA binding sites under four nucleotides (A,G,C,T) using TargetScan[24]. This prediction is facilitated by the Perl scripts of TargetScan 7.0 with default parameters. Therefore, for each mutation, we obtained two sets of miRNAs that regulate the mRNA respectively before and after the mutation. If these two sets are identical, the mutation is considered to be miRNA-synonymous; otherwise, it is considered to be miRNA-nonsynonymous.

Calculation of *i*Ka/Ks of orthologous gene pairs

Following the methodology of KaKs Calculator 3.0[7], for a pair of orthologous genes between

human and mouse, insertion/deletion mutations were eliminated after sequence alignment using MAFFT. The number of nonsynonymous sites in the coding sequence (CDS) of the human gene is denoted as *N*, while the number of synonymous sites is denoted as *S*. Next, referencing the human gene, mutations in the mouse gene were tallied with *Nd* denoting the number of nonsynonymous mutations and *Sd* denoting the number of synonymous mutations. Consequently, the conventional Ka/Ks algorithm is as follows:

$$Ka/Ks \stackrel{\text{def}}{=} \frac{\frac{Nd}{N}}{\frac{Sd}{S}}$$

We next define the *i*Ka/Ks algorithm under the framework of miRNA regulation. Among *S* synonymous sites, the number of sites simultaneously considered as miRNA-synonymous is denoted as S_m. We consider these sites to be genuinely "neutral" implying they do not partake in any natural selection process. Similarly, among *Sd* synonymous mutations, the number of mutations simultaneously considered as miRNA-synonymous is donated as Sd_m. Hence, the *i*Ka/Ks ratio is defined as follows:

$$iKa/Ks \stackrel{\text{def}}{=} \frac{\frac{Nd}{N}}{\frac{Sd_m}{S_m}}$$

In summary, the essence of the improved algorithm lies in optimizing the background substitution rate to more accurately represent the random mutation rate unaffected by natural selection. The difference of these two algorithms is shown in Figure 1. The source code of the *i*Ka/Ks algorithm and the *i*Ka/Ks ratio data of the human-mouse orthologous genes are freely available at http://www.cuilab.cn/ikaks.

Comparison of *i*Ka/Ks and Ka/Ks ratios with gene expression distance

It is generally believed that the smaller the Ka/Ks ratio of homologous genes, the closer their expression profiles are. To authenticate the superiority of the improved algorithm, we obtained the expression profile data of orthologous genes for both human and mouse from Su *et al.*'s study[25]. This dataset comprises microarray expression profiles of 2275 orthologous gene-pairs across 27 corresponding tissues in human and mouse, including heart, kidney, lung, liver, cerebellum, testis and so on. Next, we used relative abundance (RA) to standardize the data, which can avoid errors stemming from variations in probe affinity for human and mouse tissues[26]. The RA for human or mouse gene *i* expressed in tissue *j* is defined as:

$$RA_H(i,j) = S_H(i,j) \Big/ \sum_{j=1}^{n} S_H(i,j),$$

$$RA_M(i,j) = S_M(i,j) \Big/ \sum_{j=1}^{n} S_M(i,j)$$

Here, *n* is the number of tissues considered as 27 in this study. *H* indicates human and M indicates mouse. $S_H(i,j)$ and $S_M(i,j)$ are the expression signal intensities of gene *i* in human tissue *j* and mouse tissue *j*, respectively. Subsequently, the Manhattan distance is used to define the expression distance of orthologous gene *i* between human and mouse:

$$D(i) = \sum_{j=1}^{n} |RA_H(i,j) - RA_M(i,j)|$$

To validate the accuracy of the improved algorithm, we calculated the Spearman's correlation between two Ka/Ks ratios and the expression distance *D(i)*, and then performed permutation tests by random assignment for 10000 times.

Results

The distribution of Ka/Ks and *i*Ka/Ks

For each human-mouse orthologous gene pair, we first counted the number of synonymous mutations affecting miRNA binding. The distribution of genes with synonymous mutations affecting miRNA binding is shown in Figure 2A. Among all the 22,106 orthologous gene pairs, half of them have over 100 synonymous mutations that can affect miRNA binding. This result suggests a large number of synonymous mutations potentially subject to natural selection through miRNA involvement, consistent with previous findings by Wang *et al* [15]. We next calculated the conventional Ka/Ks ratio and the *i*Ka/Ks ratio for each human-mouse orthologous gene pair. As a result, globally the *i*Ka/Ks ratio is highly correlated with the conventional one (R^2=0.81, p-value=0.00). In the range where the Ka/Ks ratio is less than 0.4, their distributions are very close (D=0.0074, p-value=0.77, Kolmogorov-Smirnov test). However, as shown in Figure 2B, *i*Ka/Ks ratio is clearly distinguished from Ka/Ks for gene pairs with ratio>0.4 (D=0.16, p-value=1.17e-68, Kolmogorov-Smirnov test). It is noteworthy that among all the 22,106 orthologous gene pairs, *i*Ka/Ks exhibits 366 pairs with a ratio greater than 1, whereas the Ka/Ks only yields 11 pairs (OR=33.8, p-value=7.55e-94, Fisher's Exact test). This indicates the potential presence of significant functional differences between orthologous genes of humans and mouse.

The positive selection genes and the purifying selection genes identified by *i*Ka/Ks are consistent with the existing evidence

Among the 366 pairs of positive selection orthologous genes identified by iKa/Ks, 360 pairs

have not been identified by conventional Ka/Ks. We first examined the cellular components of these positive selection genes in human using GO enrichment analysis. As a control, we also selected the top 500 pairs of orthologous genes with the largest differences between Ka/Ks ranking and *i*Ka/Ks ranking, simultaneously restricting *i*Ka/Ks to less than 0.1, as the purifying selection genes. The top 10 cellular components of these two groups of genes are shown in Figure 3A. The extracellular region and external side of plasma membrane are the primary locales for positive selection genes, while genes under purifying selection are predominantly situated within the nucleus, chromatin, and cytoplasm. Additionally, we observed keratin filaments as the top-ranked cellular component in the positive selection genes. This arises from the presence of numerous members of the KRTAP family (keratin associated proteins) among the positive selection genes identified by *i*Ka/Ks (Figure 3B). Importantly, several researches have reported rapid divergent evolution of KRTAPs between species[27,28], which aligns with our findings. These findings illustrate that *i*Ka/Ks can identify positive selection genes as well as purifying selection genes missed by the conventional algorithms.

Furthermore, we use DisGeNET[29] to retrieve the number of human diseases associated with the genes and compared between the purifying selection genes and the positive selection genes. This result revealed that genes identified by *i*Ka/Ks undergoing purifying selection exhibited significantly more associated diseases compared to those undergoing positive selection (Figure 3C, Median value 30 vs. 23, p-value = 0.012, Wilcoxon Rank Sum test), further affirming the usefulness of *i*Ka/Ks.

_i_Ka/Ks shows stronger correlation with expression distance compared to Ka/Ks

Next, we test whether the new algorithm outperforms the conventional one using their correlation with expression distance. It is widely hypothesized that homologous genes with notable sequence differences may exhibit considerable variations in expression. For doing so, we first obtained the gene expression profiles for 2275 pairs of homologous genes across 27 tissues in human and mouse. Subsequently, we calculated their expression distance, herein defined as the relative abundance variances across distinct tissues.

We first investigated whether the new algorithm outperforms the conventional algorithm within gene sets characterized by the maximum and the minimum expression distance. As shown in Figure 4A, it is obvious that genes with smaller expression distance have significantly smaller value of *i*Ka/Ks ratio compared to Ka/Ks ratio (Median value 0.097 *vs.* 0.093, p-value=0.029, Wilcoxon Signed Rank test). While the top 20% genes with larger expression distance exhibits a significantly greater value of *i*Ka/Ks ratio compared to Ka/Ks (Median value 0.20 *vs.* 0.21, p-value=0.012, Wilcoxon Signed Rank test). These results indicates that *i*Ka/Ks can better reflect the expression distance of orthologous genes than the conventional one.

Furthermore, we investigated whether the new algorithm exhibits a stronger correlation with expression distance compared to the conventional algorithm. Given that *i*Ka/Ks ratio is globally highly correlated with Ka/Ks, here we only focused on genes with highest difference between these two algorithms. Thus, we ranked genes by their Ka/Ks values within each algorithm and selected the top *N* genes with the greatest rank differences (here *N*=50, 100, 200, 300, respectively). Subsequently, correlation analysis between the *i*Ka/Ks or Ka/Ks ratio and the expression distance was performed within these four groups. As a result, *i*Ka/Ks

showed higher correlation with the expression distance than Ka/Ks across all of the four groups (Figure 4B, Rho=0.48, 0.31, 0.28, 0.20 *vs.* 0.22, 0.057, 0.15, 0.096, p-value=4.7e-4, 1.6e-3, 8.1e-5, 4.3e-4 *vs.* 0.12, 0.58, 0.035, 0.096 for *N*=50, 100, 200, 300, respectively, Spearman's correlation). For the groups for *N*=50, 100, and 300, the conventional algorithm didn't even achieve statistical significance (with p-values of 0.12, 0.58, and 0.096, respectively). This result indicates that the new algorithm exhibits a significantly stronger correlation with expression distance than the conventional algorithm. To comprehensively test the significance of *i*Ka/Ks algorithm, we further performed permutation tests by random assignment. The result indicated that *i*Ka/Ks shows significantly higher correlation with expression distance than the conventional ratio (Supplementary Figure S1, p-value=0.0798, 0.0052, 0.0175, 0.0081 for *N*=50, 100, 200, 300, respectively. Permutation test for 10000 times). These results together suggest that the new algorithm outperforms the conventional one in this expression distance comparison.

Case studies

By several case studies, here we showed that *i*Ka/Ks algorithm is able to identify positively selected genes that were estimated to be under purifying selection by conventional Ka/Ks ratio. For example for the human-mouse orthologous transmembrane protein 72 gene *TMEM72* and *Tmem72* [30], its conventional Ka/Ks ratio is 0.21, whereas its *i*Ka/Ks is 1.13, ranking 21,903 out of 22,106 gene pairs (Supplementary Figure S2A-B). This indicates a relatively poor conservation of this homologous gene between human and mouse, and suggests positive selection during evolution and potentially substantial functional divergence. According to GENEVESTIGATOR®[31], the human *TMEM72* is specifically expressed in the

kidney and rectum, with no significant expression in the nervous system as shown in Supplementary Figure S3. In contrast, as shown in Figure S4, its mouse orthologous is highly expressed not only in kidney but also in the nervous system, including the lumbar spinal ganglion, brain microvessel and the cerebellum *etc*. To further validate this finding, we employed the cross-species functional divergence analysis proposed by Cui *et al*. [32], which defined the functional divergence score (FDS) for orthologous genes between human and mouse and is able to discover conserved functions and species-specific functions of the orthologous genes. As a result, this gene has functions of cell migration, osteoblast differentiation, and vascular remodeling in both human and mouse (Figure 5). However, it has much more specific functions in human or in mouse. It shows functions on nervous system including neuron development and neuron apoptosis only in mouse but not in human, which is consistent with our findings above. Furthermore, it shows functions of latent virus replication and antiviral immunity *etc*. only in human but not in mouse. This significant divergence in expression pattern and in functions suggest that this gene could be indeed under positive selection.

Similarly, there exist another pair of orthologous genes, *RAMP1/Ramp1*, whose *i*Ka/Ks ratio (1.05) is also much greater than Ka/Ks ratio (0.42). It was reported by NCBI that RAMP1 is a member of the single transmembrane domain protein family known as the RAMP, possessing an extracellular N-terminus and a cytoplasmic C-terminus. According to the DEGs[33], this gene is essential in human pluripotent stem cells, whereas gene knockout experiments in mice from the MGI[34] database and phenotype analysis from IMPC(www.mousephenotype.org)[35] suggest that its orthologous in mouse is non-essential. This differential essentiality further

means functional divergence of this gene between human and mouse. Indeed, the functional divergence analysis revealed that the FDS score of *RAMP1/Ramp1* is 0.89, ranking 13,979 out of total 15,505 pairs, indicating that this pair of orthologous genes is highly diverged in function between human and mouse. This result suggests that this gene may be indeed under positive selection, which also supports the *i*Ka/Ks algorithm.

The new algorithm not only can identify the homologous gene pairs with significant differences (as shown in the above two case studies) but also is able to reveal functionally conserved orthologous genes between human and mouse. For instance, the *i*Ka/Ks ratio of the *IL1B/Il1b* orthologs is 0.23, which is smaller than conventional Ka/Ks ratio (0.42). The protein encoded by this gene belongs to the interleukin-1 (IL-1) cytokine family. The IL-1 cytokine is an important mediator of the inflammatory response, and is involved in a variety of cellular activities, including cell proliferation, differentiation, and apoptosis[36,37]. According to DEGs and MGI, its orthologues in human and mouse are both essential genes. Furthermore, according to DisGeNET, it is associated with 1801 diseases. Clearly, this gene plays a crucial role in the growth and development processes of both human and mouse, exhibiting a high degree of conservation, which is consistent with the analysis results and further supports *i*Ka/Ks algorithm.

Discussion

Comparative genomics has yielded substantial insights in biology and medicine. Among several animal models, mouse stands out as the preferred model organism for studying human diseases owing to their compact size, docile disposition, swift reproductive cycle, and cost-

effectiveness in maintenance[38]. Therefore, the similarity of orthologous genes between human and mouse serves as critical information, influencing the translatability of experiments conducted in mouse models[39]. While the Ka/Ks ratio has traditionally been a widely referenced metric for measuring evolutionary rate and selective pressure in comparative or evolutionary genomics One basic assumption of the Ka/Ks algorithm is that synonymous mutations are evolutionarily neutral, which, however, has recently been found to be inaccurate, that is a large number of synonymous mutations could be not neutral but functional

Based on the above observation, this study proposes a new Ka/Ks algorithm, *i*Ka/Ks, specifically focusing on the impact of synonymous mutations on miRNA binding sites. The results showed that the algorithm can more accurately assess the functional divergence, evolutionary rate and selective pressure of the orthologous genes between human and mouse. Specifically, the new algorithm demonstrates a stronger correlation with the divergence of gene expression patterns between human and mouse.

Furthermore, evidence supports that *i*Ka/Ks rectifies previous misidentifications of genes as either conservative or non-conservative. For instance, the *TMEM72/Tmem72* pair, is previously considered to be under purifying selection, shows an *i*Ka/Ks ratio > 1, indicating substantial divergence consistent with additional evidence.

In the future, this study could be continuously improved. While the *i*Ka/Ks ratio primarily focuses on sequence-level disparities in the coding region of orthologous genes, it is essential to recognize that other factors contribute to functional divergence among these genes. Research suggests that epigenetic modifications[40], mutations in non-coding regulatory

elements[41], and the regulation of three-dimensional genome structure[42,43] also play significant roles. Future studies may aim for comprehensive analyses that integrate information from upstream and downstream regions of the entire genome, as well as structural insights, to gain a deeper understanding of the functional disparities among orthologous genes. In addition, applying *i*Ka/Ks to SNPs among human populations may help to find some important mutation sites playing important roles in human phenotype and disease.

Acknowledgements

This study was supported by the grants from the National Natural Science Foundation of China [62025102, 32301239, 81921001] and the Scientific and Technological Research Project of Xinjiang Production and Construction Corps [2023AB057].

Author Contributions

Q.C. proposed the original idea and supervised the study. J.Y. performed the study. C.C. and R.F. provided helps in implementing the new algorithm. J.Y. wrote the raw manuscript and Q.C. edited the manuscript.

References

1. Kapli, P., Yang, Z., and Telford, M.J. (2020). Phylogenetic tree building in the genomic age. Nat Rev Genet *21*, 428-444. 10.1038/s41576-020-0233-0.
2. Meadows, J.R.S., and Lindblad-Toh, K. (2017). Dissecting evolution and disease using comparative vertebrate genomics. Nat Rev Genet *18*, 624-636. 10.1038/nrg.2017.51.
3. Sivashankari, S., and Shanmughavel, P. (2007). Comparative genomics - a perspective. Bioinformation *1*, 376-378. 10.6026/97320630001376.
4. Bornstein, K., Gryan, G., Chang, E.S., Marchler-Bauer, A., and Schneider, V.A. (2023). The NIH Comparative Genomics Resource: addressing the promises and challenges of comparative genomics on human health. BMC Genomics *24*, 575. 10.1186/s12864-023-09643-4.
5. Ellegren, H. (2008). Comparative genomics and the study of evolution by natural

selection. Mol Ecol *17*, 4586-4596. 10.1111/j.1365-294X.2008.03954.x.

6. Fay, J.C., and Wu, C.I. (2003). Sequence divergence, functional constraint, and selection in protein evolution. Annu Rev Genomics Hum Genet *4*, 213-235. 10.1146/annurev.genom.4.020303.162528.
7. Zhang, Z. (2022). KaKs_Calculator 3.0: Calculating Selective Pressure on Coding and Non-coding Sequences. Genomics Proteomics Bioinformatics *20*, 536-540. 10.1016/j.gpb.2021.12.002.
8. Hurst, L.D. (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet *18*, 486. 10.1016/s0168-9525(02)02722-1.
9. Kimura, M. (1979). The neutral theory of molecular evolution. Sci Am *241*, 98-100, 102, 108 passim. 10.1038/scientificamerican1179-98.
10. Xing, Y., and Lee, C. (2006). Can RNA selection pressure distort the measurement of Ka/Ks? Gene *370*, 1-5. 10.1016/j.gene.2005.12.015.
11. Sauna, Z.E., and Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet *12*, 683-691. 10.1038/nrg3051.
12. Shen, X., Song, S., Li, C., and Zhang, J. (2022). Synonymous mutations in representative yeast genes are mostly strongly non-neutral. Nature *606*, 725-731. 10.1038/s41586-022-04823-w.
13. Chamary, J.V., Parmley, J.L., and Hurst, L.D. (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet *7*, 98-108. 10.1038/nrg1770.
14. Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., et al. (2005). Identification of hundreds of conserved and nonconserved human microRNAs. Nat Genet *37*, 766-770. 10.1038/ng1590.
15. Hurst, L.D. (2006). Preliminary assessment of the impact of microRNA-mediated regulation on coding sequence evolution in mammals. J Mol Evol *63*, 174-182. 10.1007/s00239-005-0273-2.
16. Loeb, G.B., Khan, A.A., Canner, D., Hiatt, J.B., Shendure, J., Darnell, R.B., Leslie, C.S., and Rudensky, A.Y. (2012). Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting. Mol Cell *48*, 760-770. 10.1016/j.molcel.2012.10.002.
17. Helwak, A., Kudla, G., Dudnakova, T., and Tollervey, D. (2013). Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell *153*, 654-665. 10.1016/j.cell.2013.03.043.
18. Gartner, J.J., Parker, S.C., Prickett, T.D., Dutton-Regester, K., Stitzel, M.L., Lin, J.C., Davis, S., Simhadri, V.L., Jha, S., Katagiri, N., et al. (2013). Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. Proc Natl Acad Sci U S A *110*, 13481-13486. 10.1073/pnas.1304227110.
19. Brest, P., Lapaquette, P., Souidi, M., Lebrigand, K., Cesaro, A., Vouret-Craviari, V., Mari, B., Barbry, P., Mosnier, J.F., Hebuterne, X., et al. (2011). A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. Nat Genet *43*, 242-245. 10.1038/ng.762.
20. Wang, Y., Qiu, C., and Cui, Q. (2015). A Large-Scale Analysis of the Relationship of

Synonymous SNPs Changing MicroRNA Regulation with Functionality and Disease. Int J Mol Sci *16*, 23545-23555. 10.3390/ijms161023545.

21. Martin, F.J., Amode, M.R., Aneja, A., Austine-Orimoloye, O., Azov, A.G., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., et al. (2023). Ensembl 2023. Nucleic Acids Res *51*, D933-D941. 10.1093/nar/gkac958.
22. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol *30*, 772-780. 10.1093/molbev/mst010.
23. Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. Nucleic Acids Res *47*, D155-D162. 10.1093/nar/gky1141.
24. Agarwal, V., Bell, G.W., Nam, J.W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. Elife *4*. 10.7554/eLife.05005.
25. Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A *101*, 6062-6067. 10.1073/pnas.0400782101.
26. Liao, B.Y., and Zhang, J. (2006). Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol Biol Evol *23*, 530-540. 10.1093/molbev/msj054.
27. Wu, D.D., Irwin, D.M., and Zhang, Y.P. (2008). Molecular evolution of the keratin associated protein gene family in mammals, role in the evolution of mammalian hair. BMC Evol Biol *8*, 241. 10.1186/1471-2148-8-241.
28. Wu, D.D., and Irwin, D.M. (2018). Evolution of Trichocyte Keratin Associated Proteins. Adv Exp Med Biol *1054*, 47-56. 10.1007/978-981-10-8195-8_5.
29. Pinero, J., Ramirez-Anguita, J.M., Sauch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., and Furlong, L.I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res *48*, D845-D855. 10.1093/nar/gkz1021.
30. UniProt, C. (2023). UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res *51*, D523-D531. 10.1093/nar/gkac1052.
31. Grennan, A.K. (2006). Genevestigator. Facilitating web-based gene-expression analysis. Plant Physiol *141*, 1164-1166. 10.1104/pp.104.900198.
32. Cui, C., Zhou, Y., and Cui, Q. (2021). Defining the functional divergence of orthologous genes between human and mouse in the context of miRNA regulation. Brief Bioinform *22*. 10.1093/bib/bbab253.
33. Luo, H., Lin, Y., Liu, T., Lai, F.L., Zhang, C.T., Gao, F., and Zhang, R. (2021). DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools. Nucleic Acids Res *49*, D677-D686. 10.1093/nar/gkaa917.
34. Blake, J.A., Baldarelli, R., Kadin, J.A., Richardson, J.E., Smith, C.L., Bult, C.J., and Mouse Genome Database, G. (2021). Mouse Genome Database (MGD): Knowledgebase for mouse-human comparative biology. Nucleic Acids Res *49*, D981-D987. 10.1093/nar/gkaa1083.
35. Groza, T., Gomez, F.L., Mashhadi, H.H., Munoz-Fuentes, V., Gunes, O., Wilson, R., Cacheiro, P., Frost, A., Keskivali-Bond, P., Vardal, B., et al. (2023). The International

Mouse Phenotyping Consortium: comprehensive knockout phenotyping underpinning the study of human disease. Nucleic Acids Res *51*, D1038-D1045. 10.1093/nar/gkac972.

36. Dinarello, C.A. (2018). Overview of the IL-1 family in innate inflammation and acquired immunity. Immunol Rev *281*, 8-27. 10.1111/imr.12621.
37. Dinarello, C.A. (2011). A clinical perspective of IL-1beta as the gatekeeper of inflammation. Eur J Immunol *41*, 1203-1217. 10.1002/eji.201141550.
38. Rosenthal, N., and Brown, S. (2007). The mouse ascending: perspectives for human-disease models. Nat Cell Biol *9*, 993-999. 10.1038/ncb437.
39. McGonigle, P., and Ruggeri, B. (2014). Animal models of human disease: challenges in enabling translation. Biochem Pharmacol *87*, 162-171. 10.1016/j.bcp.2013.08.006.
40. Zhao, L.Y., Song, J., Liu, Y., Song, C.X., and Yi, C. (2020). Mapping the epigenetic modifications of DNA and RNA. Protein Cell *11*, 792-808. 10.1007/s13238-020-00733-7.
41. Han, S.K., Kim, D., Lee, H., Kim, I., and Kim, S. (2018). Divergence of Noncoding Regulatory Elements Explains Gene-Phenotype Differences between Human and Mouse Orthologous Genes. Mol Biol Evol *35*, 1653-1667. 10.1093/molbev/msy056.
42. Tan, L., Xing, D., Chang, C.H., Li, H., and Xie, X.S. (2018). Three-dimensional genome structures of single diploid human cells. Science *361*, 924-928. 10.1126/science.aat5641.
43. Mohanta, T.K., Mishra, A.K., and Al-Harrasi, A. (2021). The 3D Genome: From Structure to Function. Int J Mol Sci *22*. 10.3390/ijms222111585.

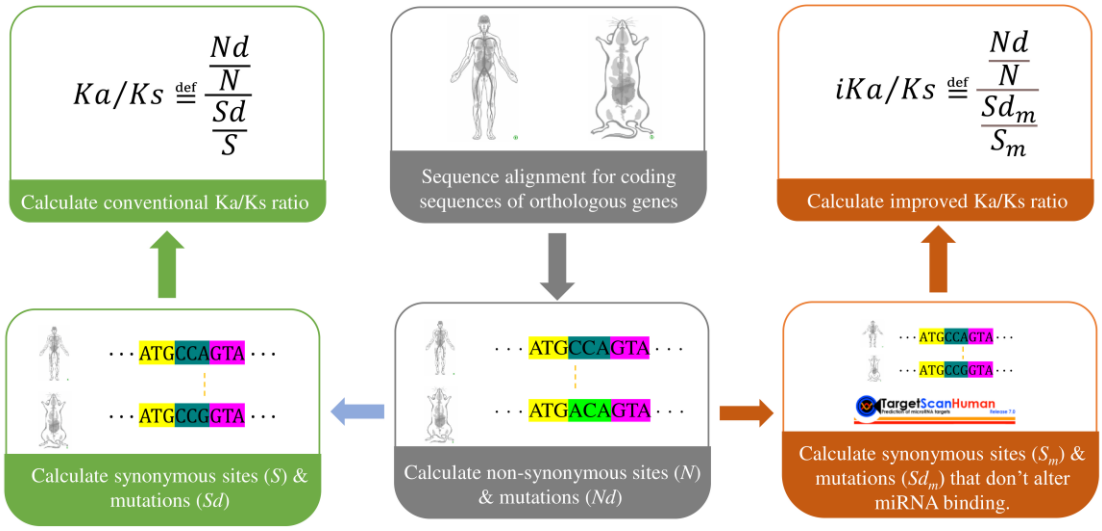

Figure 1. Workflow of the Ka/Ks algorithm and the *i*Ka/Ks algorithm.

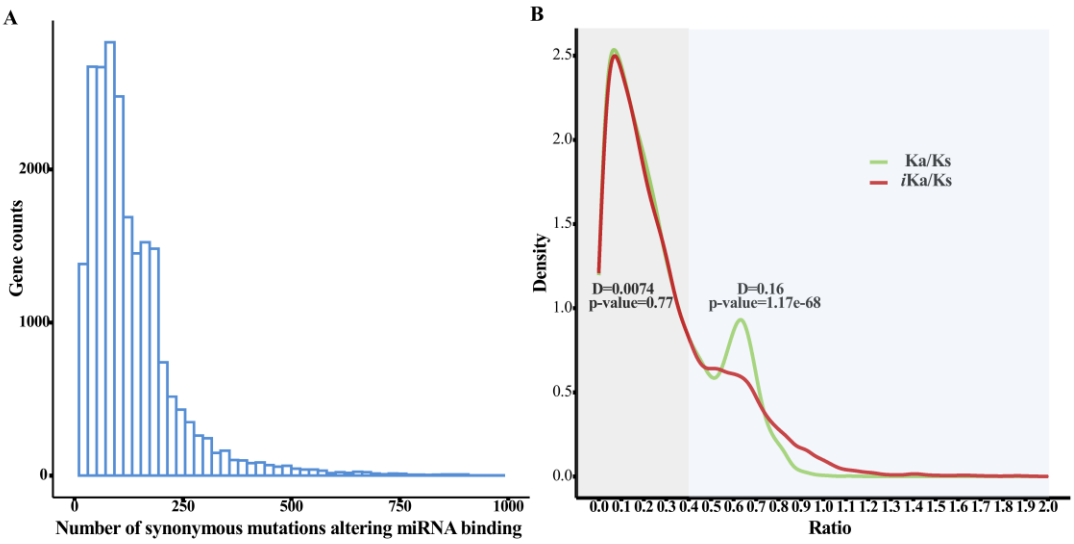

Figure 2. (A) Statistics of the number of synonymous mutations altering miRNA binding. (B) Distributions of the Ka/Ks ratio and the *i*Ka/Ks ratio. The Kolmogorov-Smirnov test is employed to assess differences between distributions. When the Ka/Ks ratio is less than 0.4, the distributions of the Ka/Ks ratio and the *i*Ka/Ks ratio are deemed highly similar (D=0.0074, p-value=0.77), whereas significant disparities between the distributions emerge when the ratio exceeds 0.4 (D=0.16, p-value=1.17e-68).

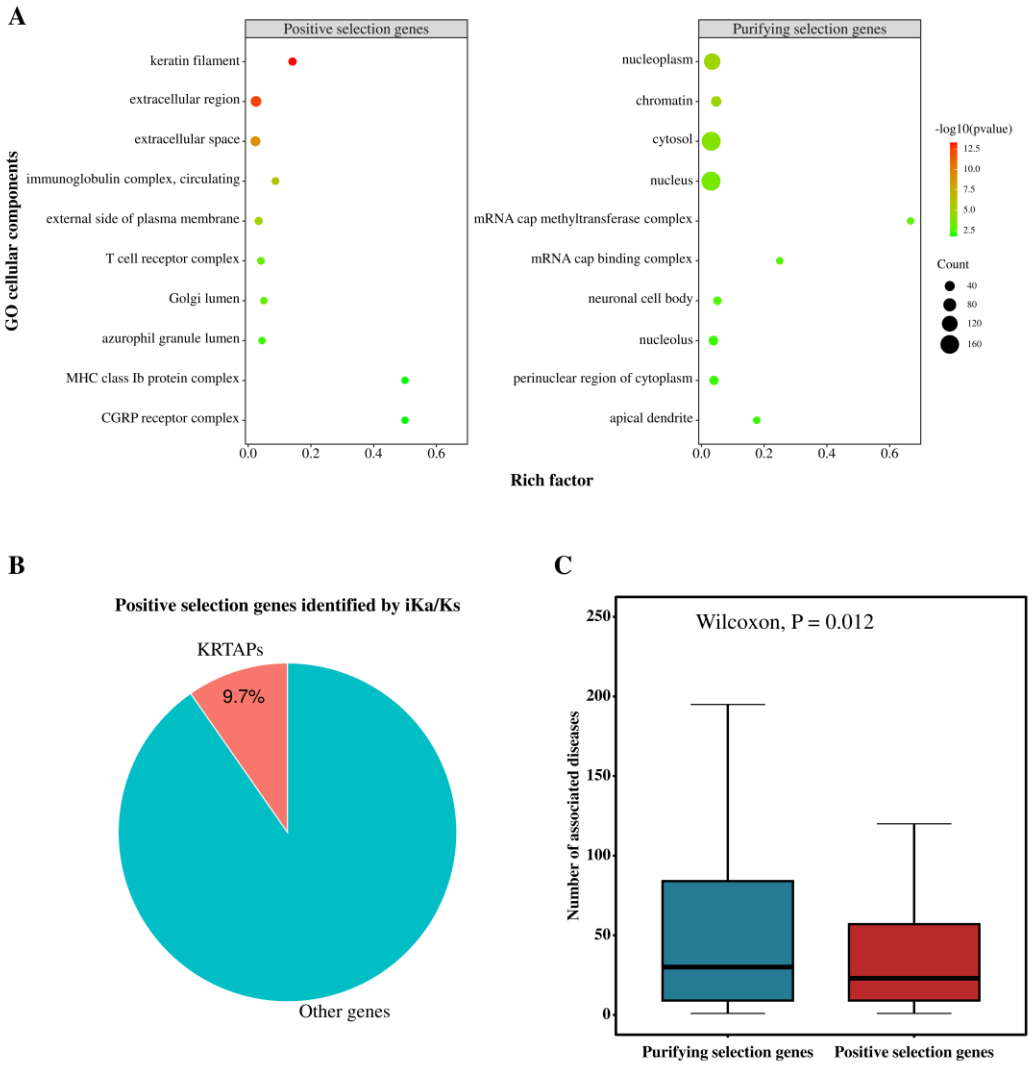

Figure 3. (A) Enrichment analysis of GO cellular components. The extracellular region and external side of plasma membrane are the primary locales for positive selection genes, while genes under purifying selection are predominantly situated within the nucleus, chromatin, and cytoplasm. Keratin filaments is the top-ranked cellular component in positive selection genes because of **(B) the presence of numerous members of KRTAPs**, which is considered as a rapid divergent evolution gene between species. (C) **Purifying selection genes exhibited significantly more associated diseases compared to positive selection genes (p-value=0.012, Wilcoxon Rank Sum test).**

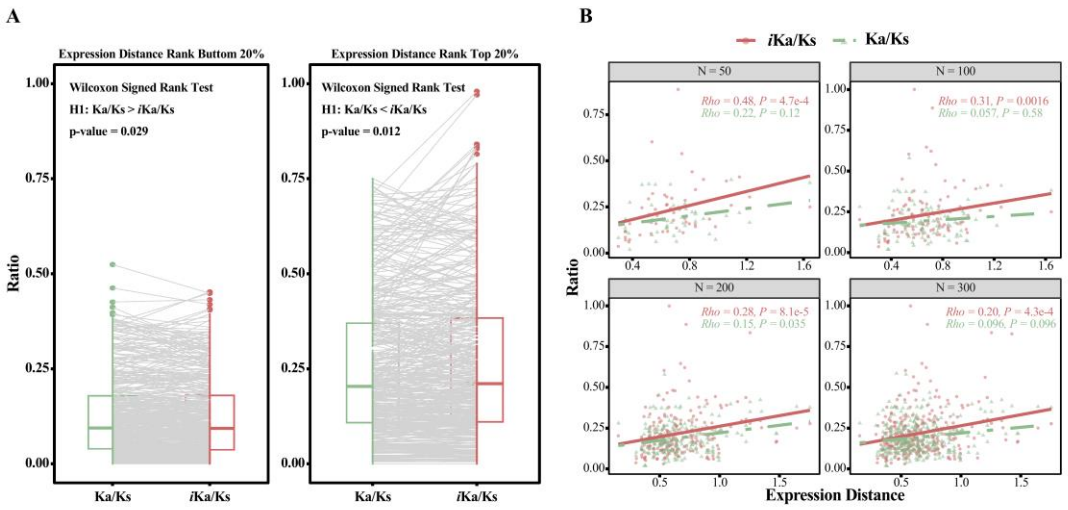

Figure 4. (A) Boxplot comparing the two algorithms grouped by the lowest and highest expression distance. In the group representing smaller expression distance, *i*Ka/Ks yields a smaller ratio (Median value 0.097 *vs.* 0.093, p-value=0.029, Wilcoxon Signed Rank test). Similarly, in the group representing larger expression distance, *i*Ka/Ks yields a larger ratio (Median value 0.20 *vs.* 0.21, p-value=0.012, Wilcoxon Signed Rank test). **(B) Correlation between *i*Ka/Ks or Ka/Ks and the expression distance.** For top *N* gene pairs with the greatest rank differences between the two algorithms, *i*Ka/Ks shows higher Rho and lower p-value (Rho=0.48, 0.31, 0.28, 0.20 *vs.* 0.22, 0.057, 0.15, 0.096; p-value=4.7e-4, 1.6e-3, 8.1e-5, 4.3e-4 *vs.* 0.12, 0.58, 0.035, 0.096 for *N*=50, 100, 200, 300, respectively).

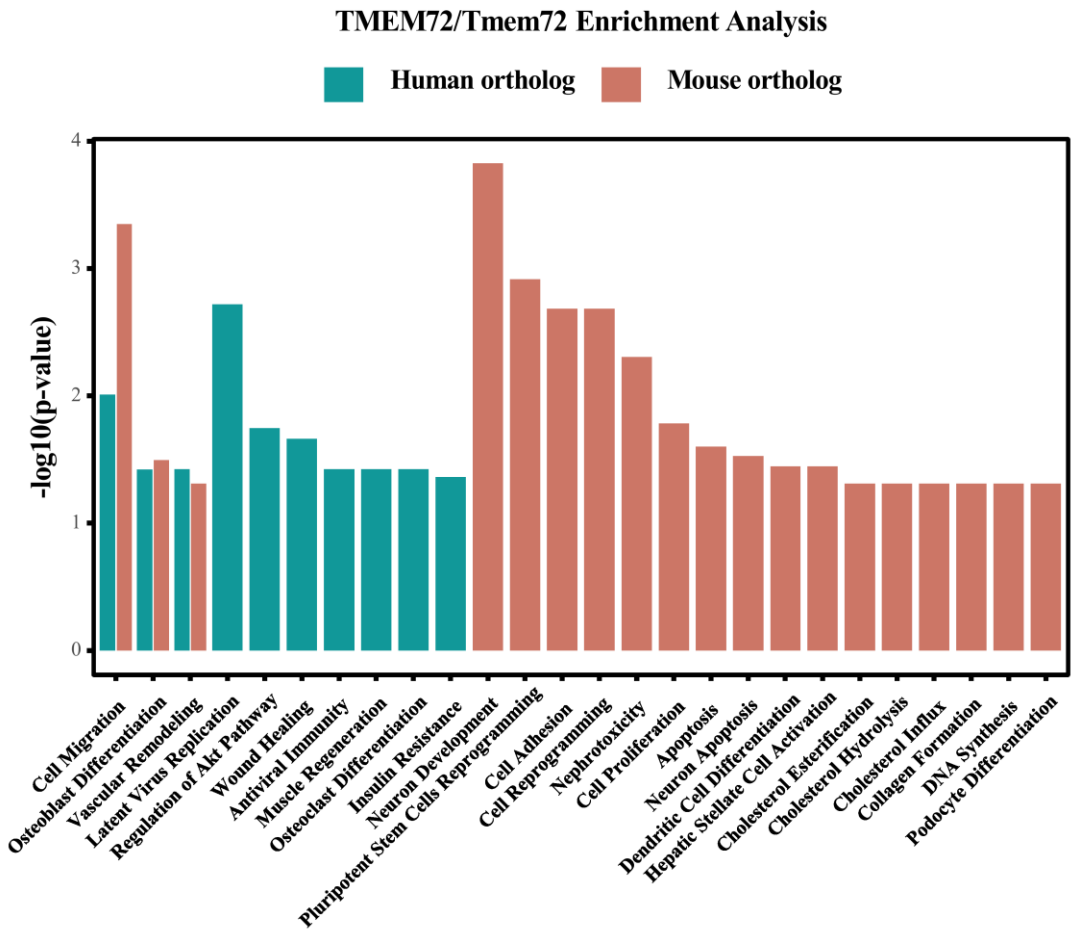

Figure 5. Functional Enrichment Analysis of TMEM72/Tmem72. This gene has functions of cell migration, osteoblast differentiation, and vascular remodeling in both human and mouse. Functions on nervous system including neuron development and neuron apoptosis are only found in mouse but not in human. And functions of latent virus replication and antiviral immunity *etc*. are only found in human but not in mouse.

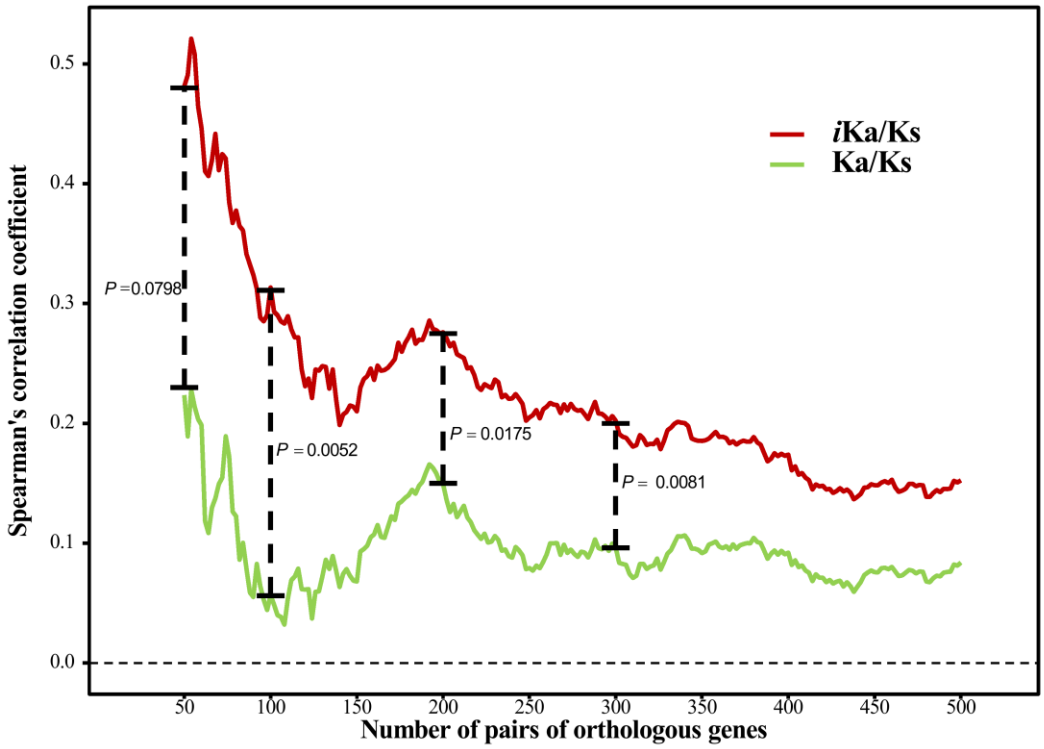

Figure S1. Correlation between *i*Ka/Ks or Ka/Ks and expression distance for $N \in$ [50,500]. The spearman's correlation coefficient with expression distance of *i*Ka/Ks is higher than that of Ka/Ks across $N \in [50,500]$. After randomizing and partitioning two sets of data, this process was repeated 10,000 times to test how often a greater difference in correlation coefficients could be obtained compared to the original arrangement (798, 52, 175, 81 times for N=50,100,200,300, respectively). This permutation test aimed to assess the significance of the superiority of the *i*Ka/Ks algorithm.

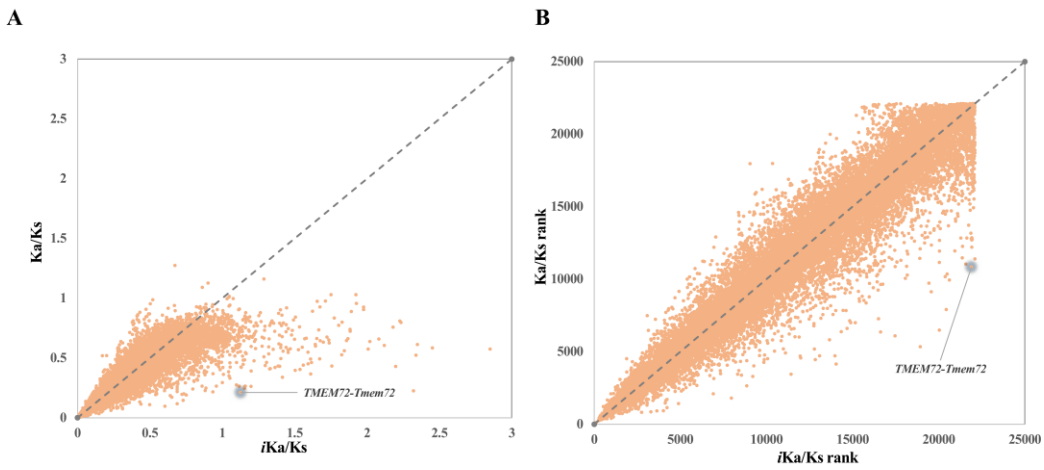

Figure S2. (A) The Ka/Ks ratio and the *i*Ka/Ks ratio of all gene pairs and the gene pair of *TMEM72-Tmem72*. (B) Rank of the Ka/Ks ratio and the *i*Ka/Ks ratio of all gene pairs and the gene pair of *TMEM72-Tmem72*.

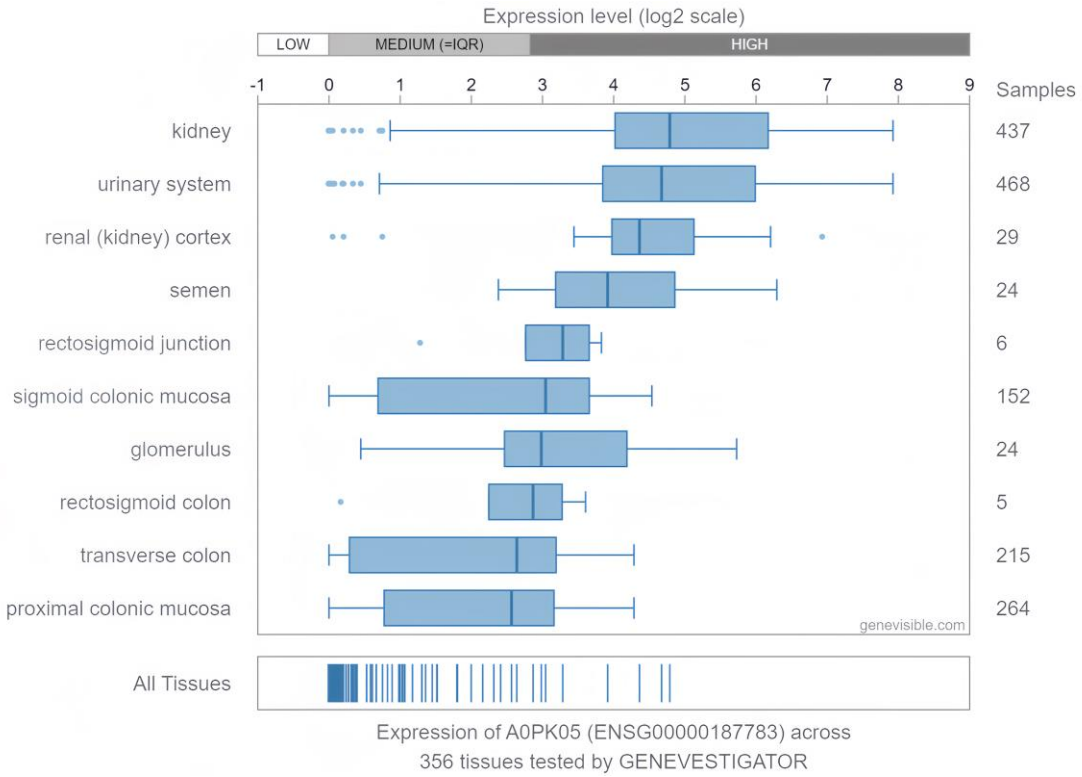

Figure S3. The expression profile of TMEM72 across 356 human tissues.

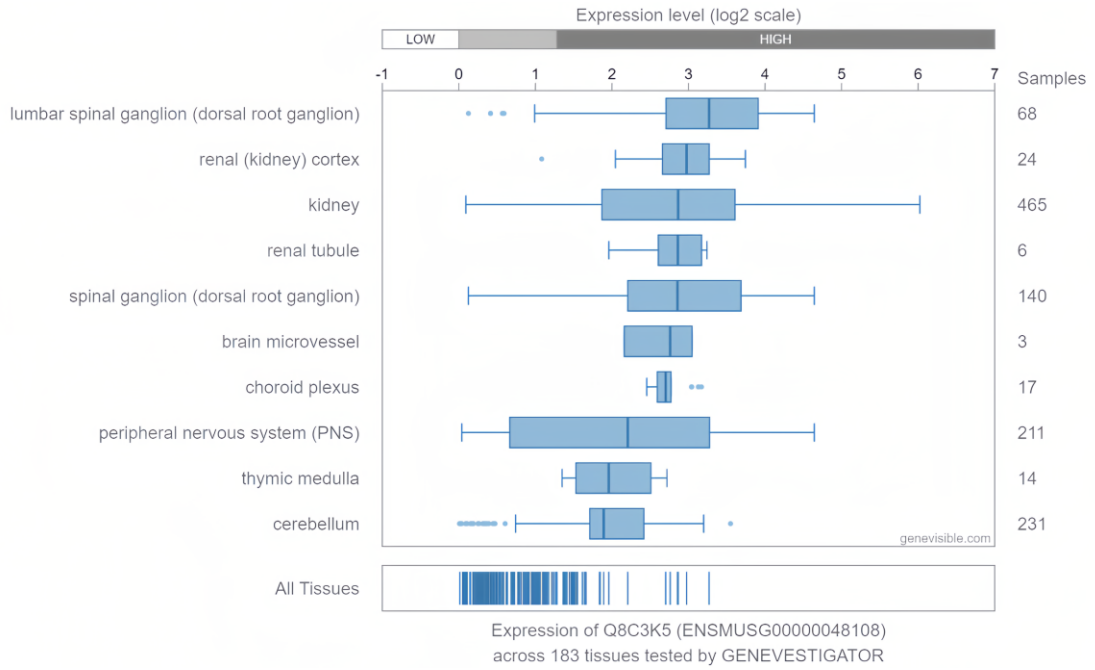

Figure S4. The expression profile of Tmem72 across 183 mouse tissues.