Using Disease-Associated Coding Sequence Variation to Investigate Functional Compensation by Human Paralogous Proteins

Sayaka Miura[1], Stephanie Tate[2] and Sudhir Kumar[1,3,4]

[1]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA. [2]School of Life Sciences, Arizona State University, Tempe, AZ, USA. [3]Department of Biology, Temple University, Philadelphia, PA, USA. [4]Center for Excellence in Genome Medicine and Research, King Abdulaziz University, Jeddah, Saudi Arabia.

ABSTRACT: Gene duplication enables the functional diversification in species. It is thought that duplicated genes may be able to compensate if the function of one of the gene copies is disrupted. This possibility is extensively debated with some studies reporting proteome-wide compensation, whereas others suggest functional compensation among only recent gene duplicates or no compensation at all. We report results from a systematic molecular evolutionary analysis to test the predictions of the functional compensation hypothesis. We contrasted the density of Mendelian disease-associated single nucleotide variants (dSNVs) in proteins with no discernable paralogs (singletons) with the dSNV density in proteins found in multigene families. Under the functional compensation hypothesis, we expected to find greater numbers of dSNVs in singletons due to the lack of any compensating partners. Our analyses produced an opposite pattern; paralogs have over 35% higher dSNV density than singletons. We found that these patterns are concordant with similar differences in the rates of amino acid evolution (ie, functional constraints), as the proteins with paralogs have evolved 33% slower than singletons. Our evolutionary constraint explanation is robust to differences in family sizes, ages (young vs. old duplicates), and degrees of amino acid sequence similarities among paralogs. Therefore, disease-associated human variation does not exhibit significant signals of functional compensation among paralogous proteins, but rather an evolutionary constraint hypothesis provides a better explanation for the observed patterns of disease-associated and neutral polymorphisms in the human genome.

KEYWORDS: functional compensation, gene duplication, Mendelian disease, single nucleotide variants, evolutionary rate

CITATION: Miura et al. Using Disease-Associated Coding Sequence Variation to Investigate Functional Compensation by Human Paralogous Proteins. *Evolutionary Bioinformatics* 2015:11 245–251 doi: 10.4137/EBO.S30594.

TYPE: Original Research

RECEIVED: June 27, 2015. **RESUBMITTED:** September 14, 2015. **ACCEPTED FOR PUBLICATION:** September 18, 2015.

ACADEMIC EDITOR: Jike Cui, Associate Editor

PEER REVIEW: Seven peer reviewers contributed to the peer review report. Reviewers' reports totaled 2,130 words, excluding any confidential comments to the academic editor.

FUNDING: This work is supported by funding from National Institutes of Health (NIH) (HG002096-12 to SK and DK098242-03 to Joel Dudley). ST was supported by an NIH training grant (R25 GM099650). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: s.kumar@temple.edu

Paper subject to independent expert blind peer review. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Published by Libertas Academica. Learn more about this journal.

Introduction

Gene duplication is an important mechanism for the origin of novelty in evolution.[1–3] When a gene is duplicated, one of the duplicate copies usually decays within a few million years due to an accumulation of deleterious mutations.[4] However, duplicates may be retained if they become functionally important to the organism.[5–7] It has been suggested that duplicate genes may be able to carry out the original gene function, which means that paralogs may compensate for each other.[8,9] Gene knockout/knockdown experiments have been conducted in multiple species to examine the degree of functional redundancy in gene families. The results suggest that the loss of function in genes with paralogs is associated with higher organismal survival than the loss of function in genes without any known paralogs (singletons), supporting the functional compensation hypothesis.[10–16] However, Liao and Zhang[17] reported that duplicates rarely compensate for each other in mice, which has been debated.[18–22] Overall, experimental data have not yet provided definitive evidence about whether paralogous genes do compensate for each other in most instances.

The predictions of functional compensation can be tested computationally by analyzing the disease-associated genetic variation in humans. These variants are currently experiencing negative selection in the human populations, which means that they constitute data of functional impact in nature. If functional compensation among gene family members is substantial, it is expected that fewer significant statistical associations between variants and disease phenotypes will be detected for proteins in multigene families than for singletons. Using this idea, Dickerson and Robertson[23] tested the predictions of functional compensation and found no difference between the proportion of singletons and paralogs implicated in diseases (2% difference), supporting the conclusions of Liao and Zhang.[17] However, they and others have suggested that recently diverged paralogs are less likely to be disease-associated than singletons and proteins with distantly related paralogs.[23–26]

These results suggest functional redundancy among young gene duplicates.

However, the abovementioned computational studies have not accounted for many potentially confounding factors. First, disease-associated single nucleotide variants (dSNVs) are found preferentially at slowly evolving amino acid positions[27]; thus, we expect to observe a higher frequency of dSNVs in more conserved proteins. This could distort comparisons between singletons and multigene family proteins if the distributions of amino acid evolutionary rates are not the same for these two classes. Second, the numbers of dSNVs found in different proteins are not expected to be the same because the numbers of amino acids in proteins vary by an order of magnitude. This means that commonly used metrics, such as the relative fractions of disease and nondisease proteins in different protein classes, are too coarse. Metrics that take into account the number of amino acids in proteins (sequence length) are necessary for more robust hypothesis testing.

In the following section, we tested the hypothesis of functional compensation by considering the abovementioned factors to better understand the genome-wide pattern of functional evolution in gene families, which is vital for understanding genome evolution and predicting disruptive effects of the mutations of proteins that have paralogs.

Results

We obtained a set of 15,485 human proteins and their homologs from 46 diverse species from the UCSC genome browser (see Material and Methods). For each protein, we also obtained a list of paralogs from the HOVERGEN database.[28] Our set of proteins is representative of the whole human gene set because about half (52%) of these proteins have at least one paralog, a fraction that is similar to the overall fraction of proteins with paralogs in the human genome (49% in HOVERGEN database[28]). For each human protein, we computed the average rate of amino acid substitution (number of substitutions per site per billion years) using the interspecific amino acid sequence alignments (see Material and Methods). Figure 1 shows the distributions of evolutionary rates in singleton and multigene family proteins. Overall, singletons are less conserved than multigene family proteins, with a ~20% mean and ~30% median difference ($P < 0.01$ by two-sample Kolmogorov–Smirnov test; Fig. 1A). Similar patterns are observed when considering paralogs belonging to small (2–5) and large (>5) multigene families ($P < 0.01$; Fig. 1B).

dSNVs in singletons and multigene families. We analyzed all available SNVs associated with Mendelian diseases in singleton and multigene family proteins. There were a total of 47,382 dSNVs in 2,589 proteins. In these data, the proportion of proteins with at least one dSNV was slightly lower (2.2%) for singletons than that of proteins with paralogs, which is consistent with the recent reports.[23,29] However, the number of dSNVs in proteins varied extensively and was found to be positively correlated with the protein length ($P < 0.05$ for multigene family and singletons; Fig. 2). This is reasonable because longer proteins should have a greater chance of accumulating random mutations and are, therefore, more likely to be classified as disease genes. Thus, we normalized the number of dSNVs by protein length to avoid any bias due to length differences in subsequent analyses.

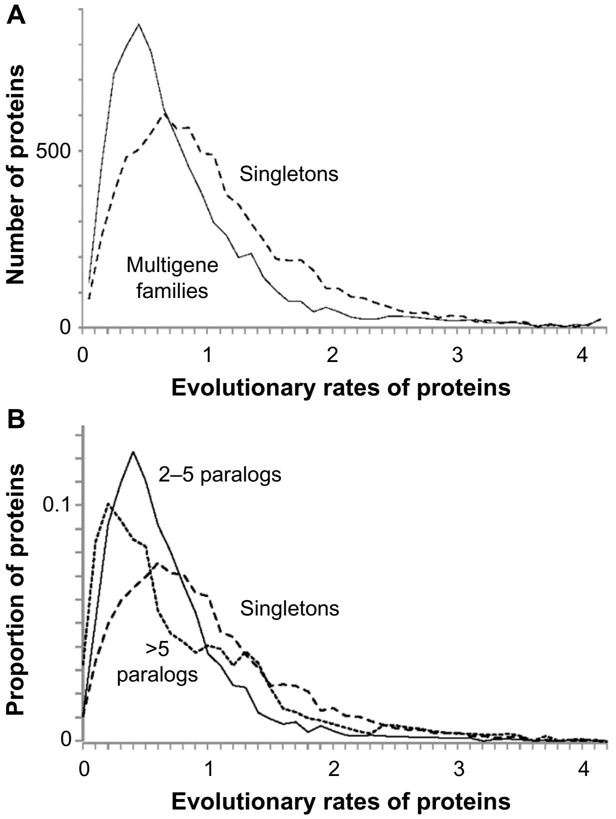

Figure 1. Distributions of evolutionary rates of singleton (broken line) and multigene family proteins (solid or dotted line). (**A**) Evolutionary rates are in the units of the number of amino acid substitutions per amino acid site per billion years. The mean and median of these distributions are 1.05 and 0.89, respectively, for singletons, and 0.80 and 0.61, respectively, for proteins in multigene families. These distributions are significantly different (two-sample Kolmogorov–Smirnov test; $P < 0.01$). (**B**) Multigene family proteins were separated into those with two to five paralogs (small family; solid line) and greater than five paralogs (large family; dotted line). The mean and median of these distributions are 0.75 and 0.60, respectively, for the proteins from the small multigene families (two to five paralogs) and 0.87 and 0.63, respectively, for the proteins from the large multigene families (greater than five paralogs). These distributions are significantly different from the distribution for singletons ($P < 0.01$).

We compared the number of dSNVs per 100 amino acid positions (dSNV density) between multigene family and singleton proteins. Multigene family proteins have 1.6 times higher density of dSNVs than detected in singleton proteins (0.66 and 0.42, respectively). We can statistically reject the null hypothesis of equal dSNV densities in singletons and multigene family proteins ($P < 0.01$). However, the direction of effect is opposite to the predictions of functional compensation from paralogous genes in multigene families, as the multigene family proteins contained significantly more dSNVs than singletons.

We tested the influence of outliers on this result by excluding all proteins with >0.5 dSNVs per amino acid. This reduced the number of proteins slightly (131 proteins were excluded), but the ratio of multigene family and singleton protein dSNV densities remained unchanged (1.6; $P < 0.01$). We, nevertheless, excluded all proteins in which the number of dSNVs per position was >0.5 in all subsequent analyses to remove the influence of proteins with unusually high dSNV density when comparing the patterns between different classes of proteins.

We also tested if the observed patterns reflect the mutations of specific amino acids (eg, arginine) that comprise a major fraction of the dataset of dSNVs (16%). Arginine codons contain a CpG dinucleotide in the first two positions and are, thus, more prone to transitional mutations, leading to amino acid variation.[30] We computed the dSNV densities using only the arginine positions in proteins and found the dSNV density in multigene family proteins to be 1.5 times greater than observed in singletons (0.09 and 0.06, respectively; $P < 0.01$). A similar pattern was observed for glycine (replacement of glycine residues occurs for 12% of dSNVs in this dataset). The dSNV density in multigene family proteins was twice than observed in singletons (0.08 and 0.04, respectively; $P < 0.01$).

Finally, we looked for the signatures of functional compensation using dSNVs that are expected to be the most severe, with the rationale that functional compensation may be easier to detect, as ameliorating severe phenotypic effects will have greater relative effect on individual fitness. We designated a dSNV to be severe if the predicted functional impact score for the variant was in the top 5% of all dSNVs (see Material and Methods). For these data, the multigene family proteins have a dSNV density 2.3 times higher than that observed for singletons (0.034 and 0.015, respectively; $P < 0.01$), which does not support the functional compensation hypothesis. Therefore, the patterns of greater abundance of dSNVs in multigene families are robust to the predicted effect sizes of dSNVs analyzed and the amino acid composition bias of the variation dataset.

Relationship of evolutionary conservation and dSNVs. We examined if protein conservation difference between singletons and multigene family proteins can explain the abovementioned pattern because it is now well established that highly conserved proteins are significantly more likely to contain dSNVs.[27,31] Because the protein evolutionary rate distributions are neither normal nor symmetrical (Fig. 1), we compared medians (0.61 and 0.89, respectively) and found a ratio of 0.69 between the multigene family and singleton proteins. The inverse of this ratio (1.5) is only slightly different from the ratio of dSNV densities (1.6). This similarity suggests that the higher rate of dSNVs in multigene family proteins is mostly explained by the degree of functional constraint on proteins in multigene families versus singleton proteins. Based on this observation, we propose the evolutionary constraints hypothesis, which posits that the differences in dSNV densities among different classes of proteins (eg, singleton vs. multigene) are primarily a result of the differences in the degree of natural selection acting upon them. If true, this would be consistent with the neutral theory of molecular evolution.[32] Evolutionary constraint hypothesis does not preclude the existence of functional compensation (among other factors) in some proteins or positions, but it does claim that differences in the intensity of purifying selection will be the primary cause of observed differences in the preponderance of SNVs in different groups of proteins.

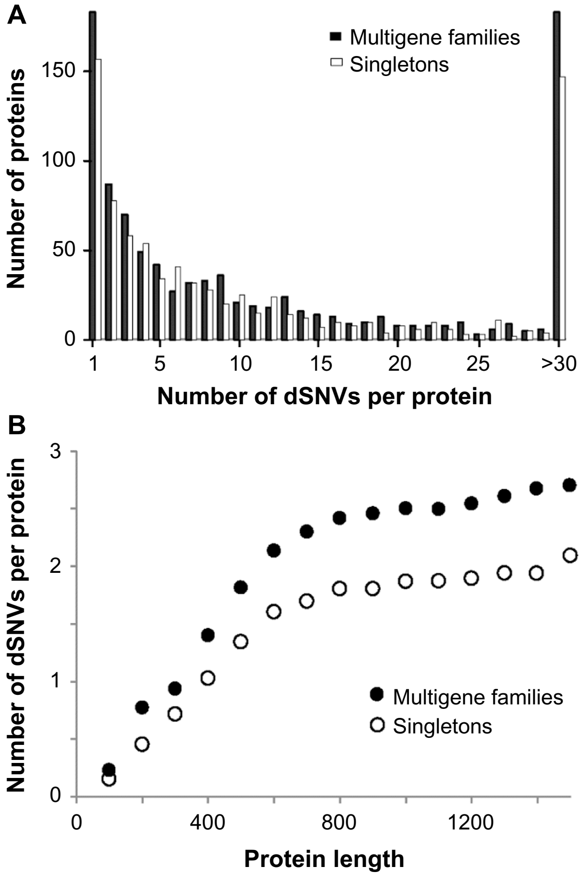

Figure 2. Distributions of the number of dSNVs. (**A**) A frequency diagram showing the number of proteins with at least one dSNV. (**B**) The average number of dSNVs per protein for proteins at different length thresholds at 100 amino acids intervals. The average number of dSNVs per protein is positively correlated with the average protein length for both multigene family (correlation = 0.005; $P < 0.01$) and singleton proteins (correlation = 0.002; $P = 0.04$).

We tested the prediction of the evolutionary constraint hypothesis in an analysis of 12,952 common neutral SNVs (nSNVs) obtained from the 1000 Genomes Project.[33] These common nSNVs are complementary in nature to dSNVs, as common nSNVs persist in the human population and have risen to moderate frequencies (>5%) because their impact on fitness is effectively neutral (opposite of dSNVs that cause disease). Therefore, if functional constraints and, thus, the conservation level of human protein sequence explain the observed differences in dSNV density, we should also observe fewer nSNVs in multigene family proteins, as these proteins evolve more slowly and are expected to be subject to more severe purifying selection.[34] Indeed, the nSNV density (number of nSNVs per 100 amino acids) in multigene family proteins was lower than that of singletons (ratio = 0.82; 0.13 and 0.16, respectively; $P < 0.01$). This ratio (0.82) is again similar to the ratio of the evolutionary rates (0.69) for these two classes of proteins. These results suggest that the occurrence of dSNVs and nSNVs in proteins is largely concordant with the degree of functional constraint on proteins, which is captured in their evolutionary rates.

Disease SNV prevalence in proteins with young and old paralogs. Next, we tested the hypothesis that functional compensation is more common in proteins with younger paralogs.[23,24] If functional compensation generally occurs only for a brief period after the gene duplication event, then the most recently diverged paralogs will provide the most powerful signal to detect functional compensation. We first identified the closest paralog for each protein within a given gene family by selecting the paralog with the smallest nucleotide divergence in their codons (third positions only). To estimate the relative antiquity of the duplicate event, we used the protein-specific human–mouse third positions in codons to normalize each closest paralog divergence across gene families (see Materials and Methods). This normalized value yields an approximate gene duplication time when it is scaled using the human–mouse divergence time (92.3 million years ago[35]). This approximation is reasonable, as third positions in codons evolve relatively neutrally and because we use divergence times primarily for identifying and sorting young paralogs for hypothesis testing.

Density of dSNV for duplicates that have diverged from their paralogs in the last 200 million years shows a tendency to increase with estimated duplicate age (Fig. 3A). The same pattern is observed for the positions of arginine and glycine and those with predicted severe functional impacts (Fig. 3B–D). Also, the dSNV densities for the youngest duplicates are lower than those for singletons (triangle in Fig. 3). We found that the evolutionary rate of proteins is negatively correlated with time since duplication, and the youngest duplicates have higher evolutionary rates than singletons (Fig. 4A). These patterns do not support the functional compensation hypothesis[23] and are consistent with our evolutionary constraint hypothesis. These trends are confirmed in the analysis of nSNV densities that showed expected complementary patterns (Fig. 4B).

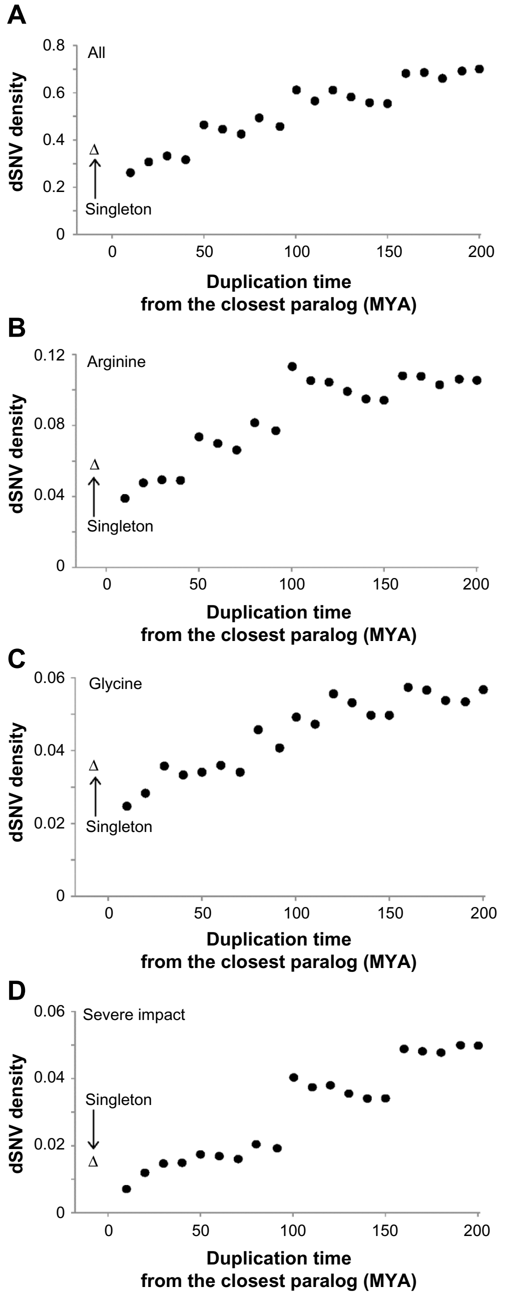

Figure 3. The dSNV density in duplicates over time. Each point shows the dSNV density of all proteins with duplication age less than or equal to a threshold time (x-axis; 10 million year intervals). The dSNV density of singletons is shown with a triangle. Panels show patterns obtained for all dSNVs (**A**), arginine dSNVs (**B**), and glycine dSNVs (**C**). Panel **D** shows patterns for dSNVs with severe impact predicted by EvoD.[46]

Disease SNV prevalence in proteins with very similar paralogs. We also tested the functional compensation hypothesis in proteins that show high amino acid sequence similarities with their paralogs, as studied by Hsiao and Vitkup.[24] We found that paralogs with the highest amino acid sequence similarities (>95%) actually have higher dSNV densities than other paralogs (0.98 vs. 0.57; $P < 0.01$). This is inconsistent with the functional compensation hypothesis but agrees with our evolutionary constraint hypothesis because the evolutionary rates were lower in paralogs with >95% similarity (0.59 and 0.78 substitutions/site/billion years; $P < 0.01$). Therefore, differences in the degree of functional constraint (measured using evolutionary rates) account for the observed patterns of dSNV densities.

Next, we compared nSNV densities in paralogs with >95% sequence similarity to those with ≤95% similarity.

For this comparison, we needed to be cognizant of the fact that variant calls are difficult when the paralogs have very similar DNA sequences.[36–39] This is the case for paralogs with >95% amino acid sequence similarity because most of these proteins also showed small divergences at the third positions in codons between paralogs (≤0.2 substitutions per site). To accommodate the variant call errors, we used proteins with ≤0.2 distance (third positions) for comparison between paralogs for two groups of proteins (225 and 69 proteins). The nSNV density was 0.30 and 0.52 for proteins that have paralogs with >95% and ≤95% sequence similarity, respectively ($P < 0.01$). The former proteins are more conserved (rate = 0.89) than the latter (rate = 1.97; $P < 0.01$), and so the result is consistent with the evolutionary constraint hypothesis.

Conclusions

In this article, we examined the functional compensation among paralogs as a general phenomenon through an analysis of disease-associated genetic variation in humans.[23–26] In contrast to expectations under the functional compensation hypothesis, we found that multigene families have a greater tendency to harbor dSNVs than singleton proteins. We proposed that differences in functional constraints (evolutionary constraint hypothesis) explain the observed pattern to a large degree. We confirmed that singleton proteins show lower functional constraint than proteins with identifiable duplicates in the genome, which explains the increased detection of disease-associated variation observed in multigene families.

Some recent theoretical and empirical studies suggest that functional compensation can lead to enhanced purifying selection and, therefore, may actually be associated with slower evolutionary rates.[14,40] Other studies indicate that the youngest duplicates are evolving under relaxed selection pressures, which would cause an increase in evolutionary rates for a few million years.[4] Such short-term and localized rate changes (faster or slower) will not have significant impact on the estimates of very long-term evolutionary rates that we have used to quantify the functional constraint. We have calculated the evolutionary rates using sequence differences in proteins that have accumulated changes for hundreds of millions of years across major groups of vertebrates. There is no evidence that pervasive functional compensation exists across the phylogenetic breadth and genomic scale reflected in our analyses. We expect our major conclusions to hold true in general, while acknowledging that functional compensation may occur in some multigene families and some amino acid positions. In summary, we suggest that there is a need to fully consider differences in the evolutionary conservation of proteins when studying the patterns of sequence variation and variant–phenotype associations.

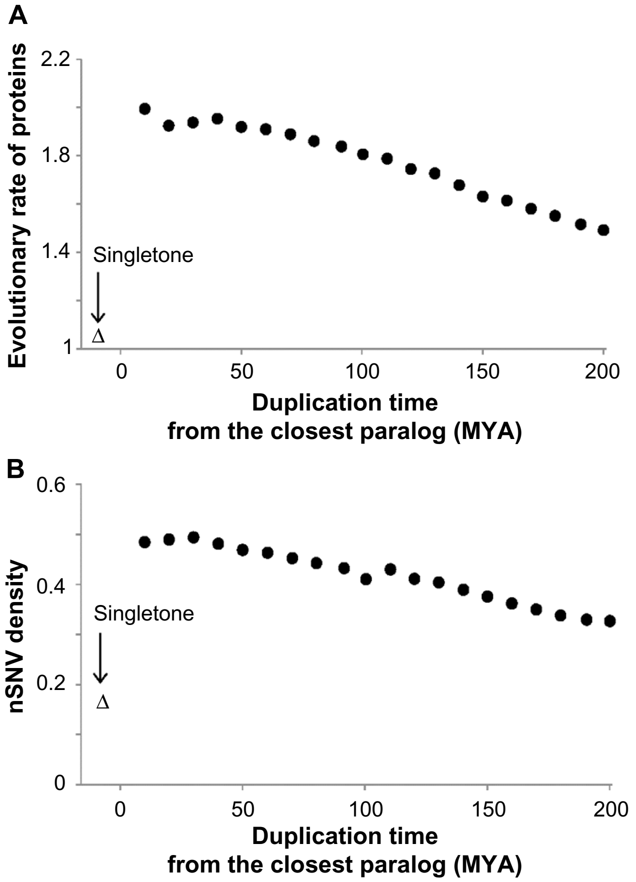

Figure 4. The average evolutionary rates (**A**) and nSNV densities (**B**) of all proteins with duplication age less than or equal to a threshold time (*x*-axis; 10 million year intervals). The decreasing trend for evolutionary rate (**A**) is opposite to that observed for dSNVs, but it is similar to that observed for nSNVs (**B**). In each panel, triangle shows the value from singletons.

Materials and Methods

Data assembly. Nonsynonymous dSNVs were obtained from the Human Gene Mutation Database (HGMD).[41] We used dSNVs associated with Mendelian diseases because they are generally caused by single mutations, which is an appropriate way to test functional compensation, as has been done before.[23,24] We excluded all SNVs associated with complex diseases whenever mutation phenotypes indicate complex disease association in the HGMD. Common nSNV data were generated by using the nonsynonymous SNV data obtained from the 1000 Genomes Project.[33] SNVs observed with a frequency >5% were assumed to be neutral (nSNVs). Nucleotide sequences of genes in the human genome and their genomic locations were obtained from the UCSC database (hg19).[42] We used protein family annotations from the HOVERGEN database.[28] RefSeq identifiers from the UCSC browser were matched with UniProt identifiers from the HOVERGEN database using the ID Mapping software.[43] Using this annotation, we made a list of paralogs for each protein. When RefSeq IDs were converted into multiple UniProt IDs, we removed those genes if they were classified into different protein families or if they were mapped to different genomic locations. As a result, we obtained 15,485 human proteins for our analyses or ~77% of human proteins in the genome (~20,000 proteins).

Estimating protein evolutionary rates. Amino acid sequence alignments for 46 diverse vertebrate species were obtained from the UCSC resource[42] to estimate the absolute site-by-site evolutionary rate as previously described.[44] For each protein, the evolutionary rate was the average of rates over all positions, which are expressed in units of substitutions per amino acid per billion years.

Testing the significance of SNV density difference between groups of proteins. To compare SNV densities between two groups (ie, between singletons and multigene family proteins), we used a z-test where the binomial variances were computed based on the proportion of the positions harboring SNVs in the groups of proteins compared. To compare the ratio of SNV densities and the ratio of evolutionary rates (this ratio was reversed for dSNV density), we conducted a bootstrap resampling test and found all differences to be highly significant because of very large sample sizes.[45]

Predicting phenotypic severities of dSNVs. For each dSNV, the evolutionary rate of the amino acid position was computed using the alignments of 46 species[44] and the impact score for EvoD prediction[46] was estimated by using myPEG.[46] The top 5% of dSNVs at ultra-conserved, well-conserved, and less-conserved positions were selected (EvoD impact scores of $\geq$88, $\geq$88, and $\geq$82, respectively). These constituted the top 5% of the most highly deleterious predicted alleles.

Identifying the closest paralog. For each protein family, codon alignments of human paralogs were built using the amino acid sequence alignment features of the MUSCLE software[47] in MEGA-CC,[48] which implements a codon alignment pipeline. Default options were used for amino acid sequence alignments (gap opening penalty = –2.9, gap extension penalty = 0, multiplier for gap open/close penalty in hydrophobic regions = 1.2, and maximum length of the diagonal = 24). To identify the closest paralog of each protein, pairwise evolutionary distances using the third positions in codons were computed between paralogs using the MEGA-CC software[48]; the maximum composite likelihood (MCL) method was used to calculate pairwise evolutionary distances.[49]

To exclude the influence of copy number variants on our analyses, we obtained the closest chimpanzee ortholog for each protein from the UCSC database (panTro2) and computed evolutionary distances at the third positions in codons (MCL) as well as among amino acids (p-distance) between the human–chimpanzee orthologous pairs. When the closest paralog pairs showed smaller evolutionary distances than those estimated between human–chimpanzee ortholog pairs, we considered those paralog pairs to be copy number variants within the human genome and removed them from further analyses. Because copy number variants are expected to have very similar nucleotide sequences, we examined paralog pairs with nucleotide divergences of <0.1 substitutions per site between the third positions in codons. Among those paralog pairs, we removed paralog pairs if chimpanzee orthologs were not found.

Estimating normalized distances between closest paralogs. For each pair of paralogs, we estimated pairwise evolutionary distance (d_p) at the third positions in codons using the MCL method in MEGA-CC.[48,49] These paralog distances were normalized using the human–mouse pairwise sequence divergence (d_{hm}) at the third positions in codons, where the mouse orthologs of the two human paralogs for each pair were obtained from the UCSC genome alignments (mm9 for mouse). d_{hm} was the average of the two human–mouse pairwise distances for the paralog pair. Note that CpG positions were identified as nucleotide C following nucleotide G, or G following C, and substitutions at CpG positions were excluded from all evolutionary distance calculations. Then, the normalized distance for a paralog pair was d_p/d_{hm}. This ratio was scaled to absolute time by multiplying it with the time of human–mouse divergence of 92.3 million years, obtained using the TimeTree resource.[50] We restricted our analysis to genes with the duplication time of <200 million years ago because the third codon position (third positions in codons) distance often exceeds 1.0 substitutions per site for older duplicates and evolutionary distances become increasingly difficult to estimate as time progresses. As our purpose is to examine young duplicates, the exclusion of old duplicates does not affect the analyses.

Acknowledgments

We thank Heather Rowe and Ravinder Kanda for many helpful comments on this manuscript.

Author Contributions

Conceived and designed the experiments: SK. Analyzed the data: SM, ST. Wrote the first draft of the manuscript: SM, SK. Contributed to the writing of the manuscript: ST. Agree with manuscript results and conclusions: SM, ST, SK. Jointly developed the structure and arguments for the paper: SK, SM. Made critical revisions and approved final version: SK, SM. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Ohno S. *Evolution by Gene Duplication*. New York: Springer-Verlag; 1970.
2. Canestro C, Albalat R, Irimia M, Garcia-Fernandez J. Impact of gene gains, losses and duplication modes on the origin and diversification of vertebrates. *Semin Cell Dev Biol*. 2013;24(2):83–94.
3. Flagel LE, Wendel JF. Gene duplication and evolutionary novelty in plants. *New Phytol*. 2009;183(3):557–64.
4. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. *Science*. 2000;290(5494):1151–5.
5. Zhang J. Evolution by gene duplication: an update. *Trends Ecol Evol*. 2003;18(6):292–8.
6. Casewell NR, Wagstaff SC, Harrison RA, Renjifo C, Wuster W. Domain loss facilitates accelerated evolution and neofunctionalization of duplicate snake venom metalloproteinase toxin genes. *Mol Biol Evol*. 2011;28(9):2637–49.
7. Deng C, Cheng CH, Ye H, He X, Chen L. Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. *Proc Natl Acad Sci U S A*. 2010;107(50):21593–8.

8. Wagner A. Gene duplications, robustness and evolutionary innovations. *Bioessays*. 2008;30(4):367–73.
9. De Smet R, Van de Peer Y. Redundancy and rewiring of genetic networks following genome-wide duplication events. *Curr Opin Plant Biol*. 2012;15(2):168–76.
10. Hannay K, Marcotte EM, Vogel C. Buffering by gene duplicates: an analysis of molecular correlates and evolutionary conservation. *BMC Genomics*. 2008;9:609.
11. Conant GC, Wagner A. Duplicate genes and robustness to transient gene knockdowns in *Caenorhabditis elegans*. *Proc Biol Sci*. 2004;271(1534):89–96.
12. Musso G, Costanzo M, Huangfu M, et al. The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. *Genome Res*. 2008;18(7):1092–9.
13. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH. Role of duplicate genes in genetic robustness against null mutations. *Nature*. 2003;421(6918):63–6.
14. Hanada K, Kuromori T, Myouga F, Toyoda T, Li WH, Shinozaki K. Evolutionary persistence of functional compensation by duplicate genes in Arabidopsis. *Genome Biol Evol*. 2009;1:409–14.
15. Vavouri T, Semple JI, Lehner B. Widespread conservation of genetic redundancy during a billion years of eukaryotic evolution. *Trends Genet*. 2008;24(10):485–8.
16. Tischler J, Lehner B, Chen N, Fraser AG. Combinatorial RNA interference in *Caenorhabditis elegans* reveals that redundancy between gene duplicates can be maintained for more than 80 million years of evolution. *Genome Biol*. 2006;7(8):R69.
17. Liao BY, Zhang J. Mouse duplicate genes are as essential as singletons. *Trends Genet*. 2007;23(8):378–81.
18. Makino T, Hokamp K, McLysaght A. The complex relationship of gene duplication and essentiality. *Trends Genet*. 2009;25(4):152–5.
19. Su Z, Gu X. Predicting the proportion of essential genes in mouse duplicates based on biased mouse knockout genes. *J Mol Evol*. 2008;67(6):705–9.
20. Chen WH, Trachana K, Lercher MJ, Bork P. Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. *Mol Biol Evol*. 2012;29(7):1703–6.
21. Plata G, Vitkup D. Genetic robustness and functional evolution of gene duplicates. *Nucleic Acids Res*. 2014;42(4):2405–14.
22. Woods S, Coghlan A, Rivers D, et al. Duplication and retention biases of essential and non-essential genes revealed by systematic knockdown analyses. *PLoS Genet*. 2013;9(5):e1003330.
23. Dickerson JE, Robertson DL. On the origins of Mendelian disease genes in man: the impact of gene duplication. *Mol Biol Evol*. 2012;29(1):61–9.
24. Hsiao T-L, Vitkup D. Role of duplicate genes in robustness against deleterious human mutations. *PLoS Genet*. 2008;4(3):e1000014.
25. Lopez-Bigas N, Ouzounis CA. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res*. 2004;32(10):3108–14.
26. Forslund K, Schreiber F, Thanintorn N, Sonnhammer EL. OrthoDisease: tracking disease gene orthologs across 100 species. *Brief Bioinform*. 2011;12(5):463–73.
27. Miller MP, Kumar S. Understanding human disease mutations through the use of interspecific genetic variation. *Hum Mol Genet*. 2001;10(21):2319–28.
28. Penel S, Arigon AM, Dufayard JF, et al. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*. 2009;10(suppl 6):S3.
29. Chen WH, Zhao XM, van Noort V, Bork P. Human monogenic disease genes have frequently functionally redundant paralogs. *PLoS Comput Biol*. 2013;9(5):e1003073.
30. de Beer TA, Laskowski RA, Parks SL, Sipos B, Goldman N, Thornton JM. Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset. *PLoS Comput Biol*. 2013;9(12):e1003382.
31. Kumar S, Dudley JT, Filipski A, Liu L. Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. *Trends Genet*. 2011;27(9):377–86.
32. Dudley JT, Kim Y, Liu L, et al. Human genomic disease variants: a neutral evolutionary explanation. *Genome Res*. 2012;22(8):1383–94.
33. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56–65.
34. Subramanian S, Kumar S. Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. *BMC Genomics*. 2006;7:306.
35. Hedges SB, Dudley J, Kumar S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*. 2006;22(23):2971–2.
36. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*. 2012;13(1):36–46.
37. Nakken S, Rodland EA, Rognes T, Hovig E. Large-scale inference of the point mutational spectrum in human segmental duplications. *BMC Genomics*. 2009;10:43.
38. Cheung J, Estivill X, Khaja R, et al. Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol*. 2003;4(4):R25.
39. Estivill X, Cheung J, Pujana MA, Nakabayashi K, Scherer SW, Tsui LC. Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. *Hum Mol Genet*. 2002;11(17):1987–95.
40. Bozorgmehr JE. The effect of functional compensation among duplicate genes can constrain their evolutionary divergence. *J Genet*. 2012;91(1):1–8.
41. Stenson PD, Ball EV, Mort M, Phillips AD, Shaw K, Cooper DN. The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Curr Protoc Bioinformatics*. 2012;Chapter 1:Unit1.13.
42. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res*. 2002;12(6):996–1006.
43. The UniProt Consortium. Reorganizing the protein space at the universal protein resource (UniProt). *Nucleic Acids Res*. 2012;40(Database issue):D71–5.
44. Kumar S, Suleski MP, Markov GJ, Lawrence S, Marco A, Filipski AJ. Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and damaging personal amino acid mutations. *Genome Res*. 2009;19(9):1562–9.
45. Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K. Statistics and truth in phylogenomics. *Mol Biol Evol*. 2012;29(2):457–72.
46. Kumar S, Sanderford M, Gray VE, Ye J, Liu L. Evolutionary diagnosis method for variants in personal exomes. *Nat Methods*. 2012;9(9):855–6.
47. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792–7.
48. Kumar S, Stecher G, Peterson D, Tamura K. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics*. 2012;28(20):2685–6.
49. Tamura K, Nei M, Kumar S. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci U S A*. 2004;101(30):11030–5.
50. Kumar S, Hedges SB. TimeTree2: species divergence times on the iPhone. *Bioinformatics*. 2011;27(14):2023–4.