Long-term survival of duplicate genes despite absence of subfunctionalized expression.

Xun Lan[1,*], Jonathan K. Pritchard[1,2,3*]

[1]Department of Genetics, Stanford University

[2]Department of Biology, Stanford University

[3]Howard Hughes Medical Institute, Stanford University

[*]Correspondence: xlan@stanford.edu, pritch@stanford.edu.

April 15, 2015

Abstract

Gene duplication is a fundamental process in genome evolution. However, young duplicates are frequently degraded into pseudogenes by loss-of-function mutations. One standard model proposes that the main path for duplicate genes to avoid mutational destruction is by rapidly evolving subfunctionalized expression profiles. We examined this hypothesis using RNA-seq data from 46 human tissues. Surprisingly, we find that sub- or neofunctionalization of expression evolves very slowly, and is rare among duplications that arose within the placental mammals. Most mammalian duplicates are located in tandem and have highly correlated expression profiles, likely due to shared regulation, thus impeding subfunctionalization. Moreover, we also find that a large fraction of duplicate gene pairs exhibit a striking asymmetric pattern in which one gene has consistently higher expression. These asymmetrically expressed duplicates (AEDs) may persist for tens of millions of years, even though the lower-expressed copies tend to evolve under reduced selective constraint and are associated with fewer human diseases than their duplicate partners. We suggest that dosage-sharing of expression, rather than subfunctionalization, is more likely to be the initial factor enabling survival of duplicate gene pairs.

Main Text

Gene duplications are a major source of new genes, and ultimately of new biological functions (*1–3*). Gene duplication is likely the primary mechanism by which new genes are born (*4*).

However, the evolutionary forces governing the initial spread and persistence of young duplicates remain controversial (*5*). New duplicates are usually functionally redundant and thus susceptible to loss-of-function mutations that degrade one of the copies into a pseudogene. The average half-life of new duplicates has been estimated at just 4 million years (*6*).

There has been a great deal of work to understand why many duplicate pairs do survive over long evolutionary timescales (*5*). These models generally assume that long-lived duplicates must evolve distinct functions to avoid mutational degradation, either by neofunctionalization (in which one copy gains new functions) or subfunctionalization (the copies divide the ancestral functions between them). One influential model known as Duplication-Degeneration-Complementation (DDC) proposes that complementary degeneration of regulatory elements may lead to the copies being expressed in different tissues, such that both copies are required to provide the overall expression of the ancestral gene (*7*). Similarly, neofunctionalization of expression could lead to one gene copy gaining function in a tissue where the parent gene was not expressed. Functional divergence may also occur at the protein level (*8*), but it is generally thought that divergence usually starts through changes in regulation (*1, 9*). Several empirical studies have measured functional redundancy of duplicates, but overall patterns and conclusions are inconsistent across organisms and approaches (*10–16*). One study of single and double knockouts of yeast duplicates reported surprisingly high levels of apparent redundancy even among old duplicate pairs (*10*).

We therefore set out to test whether modern gene expression data from many tissues in human and mouse support the standard model of duplicate preservation by subfunctionalization of expression. Based on theoretical models and previous literature, we expected that–aside from the youngest duplicates–most duplicate pairs would be functionally distinct, and that the primary mechanism for this would be through divergent expression profiles. In particular, the sub- and neofunctionalization models suggest that, for each duplicate gene, there should be at least one tissue where that gene is more highly expressed than its partner. To test this prediction, we analyzed RNA-seq data from ten individuals for each of 46 diverse human tissues collected by the GTEx Project (*17*), and replicated our general conclusions using RNA-seq from 26 diverse mouse tissues (*18*).

We first developed a computational pipeline for identifying duplicate gene pairs in the human genome (Supp. Inf. Section 2). After excluding annotated pseudogenes, we identified 1,444 high-confidence reciprocal best-hit duplicate gene pairs with $>$80% alignable coding sequence and $>$50% average sequence identity. We used synonymous divergence d_S as a proxy for divergence time, while noting that divergence of gene pairs may be downwardly biased due to nonallelic homologous gene conversion in young duplicates (*19*). We estimate that d_S for duplicates that arose at the time of the human-mouse split averages $\sim$0.4 and that most pairs with d_S $>\sim$0.7 predate the origin of the placental mammals (Figs. S3, S4).

As expected, there is a peak of very young gene pairs dating to within the apes ($d_S < 0.1$; Fig. 1A). This peak reflects the fact that only a small fraction of duplicates survive long-term (*6*). Most of the 621 pairs with $d_S < 0.7$ are physically close in the genome, likely because they were generated by segmental duplications. Older duplicates have often been separated by genomic rearrangements, although this is a very slow process. We estimate that 6% of the identified duplicates arose from retrotransposition (Fig. 1A; Supp. Inf. Section 5).

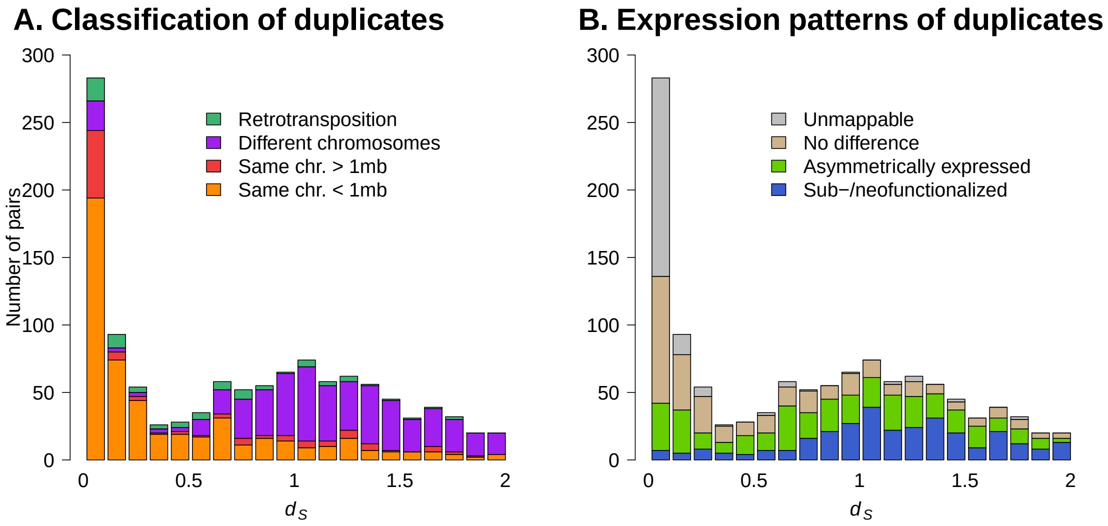

Figure 1: **Properties of duplicate gene pairs. A.** Numbers of pairs for different values of d_S, showing that most young pairs are nearby in the genome. **B.** Classification of gene pairs by expression patterns; note that sub-/neofunctionalization is rare except among older gene pairs. "Unmappable" indicates that RNA-seq reads could not be uniquely mapped to both genes. For context, note that duplicates arising at the human-mouse split would have d_S ~0.4.

We next considered GTEx RNA-seq data from 46 tissues. Accurate measurement of expression in gene duplicates can be challenging because some RNA-seq reads may map equally well to both gene copies. There are also cases where reads from one gene copy map better than reads from the other copy, due to differential homology with other genomic locations. To overcome these challenges, we developed a new method specifically for estimating the expression levels of duplicate genes (Supp. Inf. Section 3). In brief, we identified paralogous positions within each duplicate pair for which reads from both copies would map uniquely to the correct gene. Only these positions were used for estimating expression ratios. This approach is analogous to methods for measuring allele-specific expression (*20*). These strict criteria mean that some very young genes are excluded from our expression analyses as unmappable but, for the remaining genes, simulations show that our method yields highly accurate, unbiased estimates of expression ratios (Fig. S1).

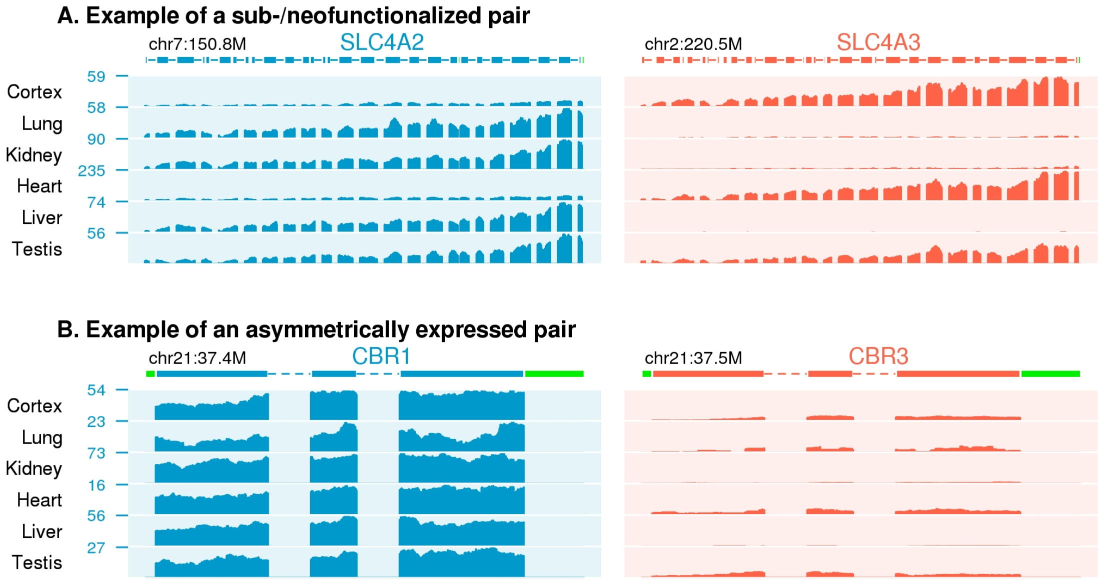

Figure 2: **Expression of duplicate genes in representative tissues. A.** A gene pair with an expression profile consistent with sub- or neofunctionalization: i.e., each gene is significantly more highly expressed than the other in at least one tissue. **B.** An asymmetrically expressed gene pair. Notice that expression of CBR1 exceeds expression CBR3 in all tissues. Introns have been shortened for display purposes. The Y-axis shows read depth per billion mapped reads. Green regions in the gene models are unmappable.

This new read mapping pipeline allowed us to classify gene pairs into categories based on their co-expression patterns (Supp. Inf. Sections 3, 6). We classified a gene pair as being *potentially sub-/neofunctionalized* if both genes have significantly higher expression (at least 2-fold difference and $p<0.001$) in at least one tissue each.

Analysis of the RNA-seq data shows that surprisingly few duplicate pairs show any evidence of sub-/neofunctionalization of expression (Fig. 1B; example in Fig. 2A). Moreover, most gene pairs that do show such patterns are very old, dating to before the emergence of the placental mammals: for duplicates with $d_S < 0.7$, just 10.7% of duplicates are classified as potentially sub-/neofunctionalized in expression. Given that even modest variation in expression profiles across tissues would meet our criteria for subfunctionalization, the fraction of truly subfunctionalized duplicates is probably even lower.

Of course some additional duplicates might show evidence for subfunctionalization in a tissue not measured by GTEx. However it is unlikely that such cases would dramatically change the overall picture: We show below that genes with evidence for subfunctionalization in our data are significantly more conserved, and have a higher burden of gene-specific diseases compared to duplicates without evidence for subfunctionalization. This argues strongly that our data are not simply an artifact of misclassification. We also find very similar patterns in a mouse dataset with better representation of fetal tissues (*18*) (Fig. S11). We wondered whether gene pairs with higher tissue specificity might be more subfunctionalized (as they may have more tissue-specific enhancers) but this is not the case (Fig. S10). Finally, we hypothesized that subfunctionalization might instead occur through differential splicing of exons (*21*); however we found little evidence for this among mammalian duplicates (Fig. S13, Supp. Inf. Section 9).

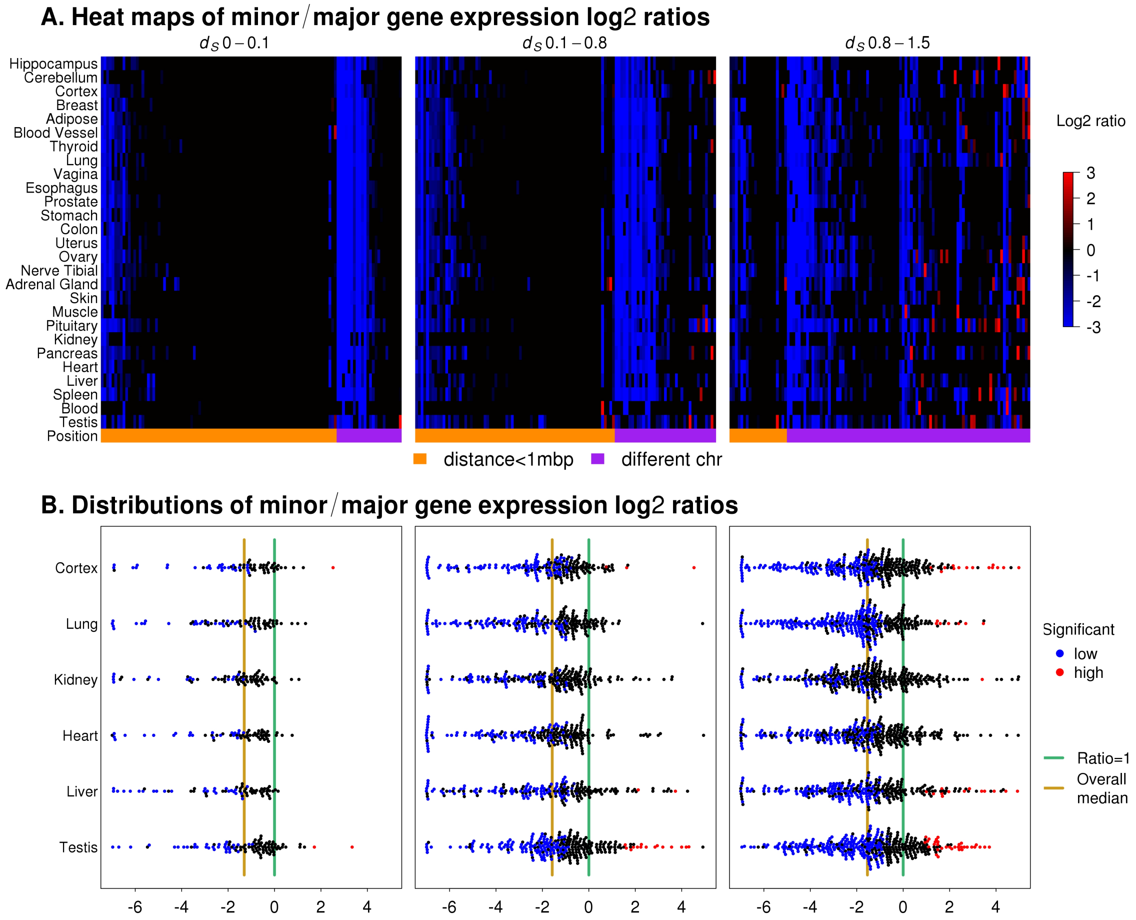

Figure 3: **Gene expression ratios for duplicate gene pairs.** **A.** Heat maps of expression ratios for a representative sample of duplicate gene pairs, at 3 levels of synonymous divergence, d_S. For each duplicate pair (plotted in columns) the ratios show the tissue-specific expression level of the gene with lower global expression relative to its duplicate. Blue indicates significantly lower expression of the minor gene in a particular tissue; red indicates significantly higher expression of the minor gene (p<.001 for both cases). Black indicates no significant difference. **B.** Distributions of expression ratios for duplicate gene pairs in representative tissues. Labeling same as in A. Notice that for most gene pairs, the minor gene has consistently lower expression than the major gene, with few clear cases of subfunctionalization (i.e., mix of red/blue) except for the most diverged gene pairs. Testis is often an outlier, enriched with sub- or neofunctionalized genes (*16*).

Surprisingly, many gene pairs instead show an unexpected asymmetric pattern of gene expression, in which one gene has consistently higher expression. For each pair, we classified the gene with higher overall expression as the "major" gene, and its partner as the "minor" gene. On average, minor genes are expressed at 40% of the level of major genes. Moreover, a large fraction of gene pairs show particularly strong asymmetry; see Fig. 2B and Fig. 3 for examples. We classified a gene pair as an *asymmetrically expressed duplicate (AED)* if the major gene was significantly more highly expressed ($p<.001$) in at least 1/3 of tissues where either gene is expressed, and not lower expressed than the minor gene in any tissue. The remaining duplicates were classified as *no difference* pairs, though many show weaker levels of asymmetry.

Among duplicates that arose within the placental mammals, AEDs are much more common than potentially subfunctionalized pairs: 36.7% of duplicates with $d_S < 0.7$, compared to just 10.7% of potentially sub-/neofunctionalized pairs. Most remaining gene pairs show little or no difference in expression within tissues (Fig. 1B; black in Fig. 3).

To learn whether AEDs are a functionally meaningful category, we examined the numbers of known diseases (*22*) associated with different types of duplicate pairs (Fig. 4A). As might be expected, AED minor genes are significantly less likely than major genes to be associated with diseases (32% vs 46%; $p=9\times10^{-5}$). Across all duplicates, there is a strong effect that the lower the expression of the minor gene compared to the major gene, the lower the disease burden of the minor gene ($p=4\times10^{-10}$, controlling for d_S; see Supp. Inf. Section 8 for details) (Table S2). In contrast, the extent of subfunctionalization is highly *positively* correlated with the number of gene-specific diseases ($p=1\times10^{-9}$) (Table S3).

AEDs are thus somewhat mysterious: why should a large class of duplicates with broadly reduced expression be maintained in the genome? Are these genes functionally constrained, or simply destined for mutational oblivion? We measured the strength of sequence conservation on major genes compared to minor genes of AEDs. Minor genes do show clear evidence of functional constraint: 97% of minor genes have d_N/d_S<1, which is a hallmark of protein-coding constraint (Fig. 4B). Nonetheless, minor genes evolve under relaxed selective constraint relative to major genes, both between species (Figs. 4B, S14) and within the human population, where minor genes have a higher rate of common missense and nonsense variants compared to major genes (Figs. 4C, S16, S17).

An alternative hypothesis for the preservation of duplicates is that they can become non-redundant by sharing the required dosage of gene expression (*25*). This model suggests that duplicates should rapidly evolve reduced expression, such that the summed expression of the duplicates is close to that of the parent gene. Subsequently, loss of either gene is deleterious because it leads to a deficit of expression.

To evaluate this, we analyzed the expression of human duplicates that arose since the human-macaque split using RNA-seq data from 6 tissues in human and macaque (*24*) (Fig. 4D, Supp. Inf. Section 7). Indeed, there is a very clear signal that both copies tend to evolve reduced expression, such that the median summed expression of the human duplicates is very close to the expression of the singleton orthologs in macaque (median expression ratio 1.11; this is significantly less than the 2:1 expression ratio expected based on copy number, p=7.6×10^{-6}). Thus, our data suggest a model of duplicate preservation by dosage sharing, rather than subfunctionalization.

Finally, to understand why subfunctionalization evolves so slowly, we explored the

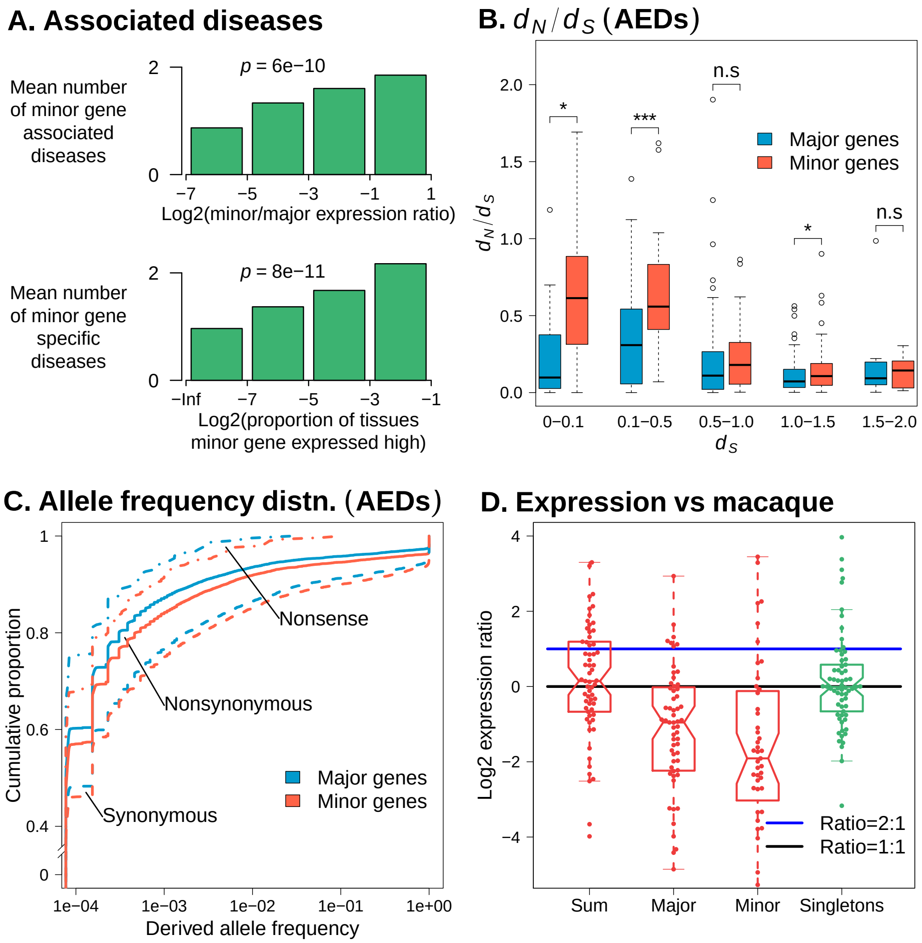

Figure 4: **Functional properties of duplicates. A.** Disease burden of minor genes as a function of overall expression asymmetry (top) and a measure of subfunctionalization (bottom). **B.** d_N/d_S for major and minor genes at AEDs, stratified by d_S. **C.** Site frequency spectra for synonymous, nonsynonymous, and nonsense mutations, plotted for major and minor genes at AEDs, respectively. The plots show cumulative derived allele frequencies at segregating sites. The lines that climb more steeply (major for each variant type) have a higher fraction of rare variants, indicating stronger selective constraint. Data from (*23*). **D.** Expression levels of young duplicates compared to their macaque orthologs in 6 tissues (*24*), for human duplicates that are single-copy genes in macaque. “Sum” shows the summed expression of both duplicates, relative to expression of the macaque orthologs in the same tissues; the “major” and “minor” data show equivalent ratios for the higher and lower expressed genes in each duplicate pair. Each tissue-gene expression ratio is plotted as a separate data point. The green data show results for a random set of singleton orthologs.

genomic features that lead to divergent expression profiles of duplicate genes. Controlling for d_S, the most important predictor of sub-/neofunctionalization is that the duplicates are found on different chromosomes (Fig. 5). Most preserved duplicate pairs arise as segmental duplications (*2*) and are found close together in the genome: 87% of gene pairs with d_S <0.1 are on the same chromosome (Fig. 1A). The duplicates may subsequently become separated as the result of chromosomal rearrangements (*26*), however this is a slow process–it is not until d_S=0.6 that half of gene duplicates are on different chromosomes.

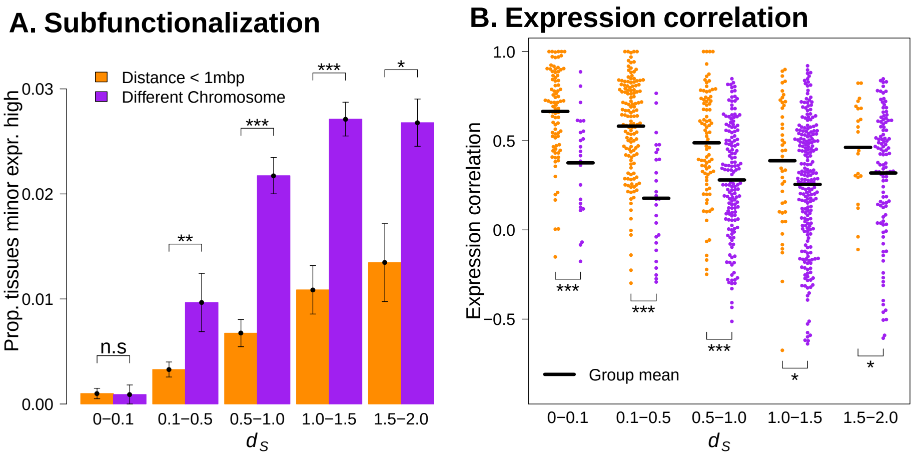

Figure 5: **Expression divergence as a function of genomic distance between duplicates.** **A.** Divergent expression profiles between duplicates: Proportions of tissues in which the gene with lower overall expression (minor gene) is more highly expressed than the major gene. **B.** Distributions of the correlation coefficients of duplicate pairs, across tissues. Pairs within 1MB (orange) tend to be more correlated than pairs on different chromosomes (purple).

Genomic separation of gene duplicates seems to be a major factor enabling expression divergence. For all values of d_S, duplicates that are separated in the genome are much more likely to be sub-/neofunctionalized (Fig. 5A) and to have less correlated expression (Figs. 5B, S22). When the duplicates are on the same chromosome, there is also a significant, though weaker, effect of the distance between the genes on expression correlation (p=0.002, Fig. S21). These effects of physical linkage may arise because tandem duplicates share regulatory elements, or are coordinately regulated within the same topological chromatin domains (*27, 28*). Separated duplicates are also more likely to evolve rapidly in protein space, as measured by high d_N/d_S (*29*), consistent with a model in which translocation enables decoupling of gene regulation, followed by subfunctionalization or relaxation of constraint.

In summary, the prevailing model for the evolution of gene duplicates holds that, to survive, duplicates must achieve non-redundant functions, and that this usually occurs by partitioning the expression space (*1,7*). However, we report here that sub-/neofunctionalization of expression occurs extremely slowly, and generally does not happen until the duplicates are separated by genomic rearrangements. Thus, in most cases long-term survival must rely on other factors. Some genes may acquire new protein functions (*8, 30, 31*), though this is likely to be a slow process in most cases; or they may function in RNA regulation (*32*).

Alternatively, duplicates may be preserved by dosage sharing (*25*). We propose that following duplication the expression levels of a gene pair evolve so that their combined expression matches the optimal level. Subsequently, the relative expression levels of the two genes evolve as a random walk, but do so slowly (*33*) due to constraint on their combined expression. If expression happens to become asymmetric, this reduces functional constraint on the minor gene. Subsequent accumulation of missense mutations in the minor gene may provide weak selective pressure to eventually eliminate expression of this gene, or may free the minor gene to evolve new functions.

References

1. Gavin C Conant and Kenneth H Wolfe. Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics*, 9(12):938–950, 2008.

2. Henrik Kaessmann. Origins, evolution, and phenotypic impact of new genes. *Genome Research*, 20(10):1313–1326, 2010.

3. Sidi Chen, Benjamin H Krinsky, and Manyuan Long. New genes as drivers of phenotypic evolution. *Nature Reviews Genetics*, 14(9):645–660, 2013.

4. Roy J Britten. Almost all human genes resulted from ancient duplication. *Proceedings of the National Academy of Sciences*, 103(50):19027–19032, 2006.

5. Hideki Innan and Fyodor Kondrashov. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, 11(2):97–108, 2010.

6. Michael Lynch and John S Conery. The evolutionary demography of duplicate genes. *Journal of Structural and Functional Genomics*, 3(1-4):35–44, 2003.

7. Allan Force, Michael Lynch, F Bryan Pickett, Angel Amores, Yi-lin Yan, and John Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545, 1999.

8. Christopher R Baker, Victor Hanson-Smith, and Alexander D Johnson. Following gene duplication, paralog interference constrains transcriptional circuit evolution. *Science*, 342(6154):104–108, 2013.

9. Ilan Wapinski, Avi Pfeffer, Nir Friedman, and Aviv Regev. Natural history and evolutionary principles of gene duplication in fungi. *Nature*, 449(7158):54–61, 2007.

10. E Jedediah Dean, Jerel C Davis, Ronald W Davis, and Dmitri A Petrov. Pervasive and persistent redundancy among duplicated genes in yeast. *PLoS genetics*, 4(7):e1000113, 2008.

11. Kousuke Hanada, Takashi Kuromori, Fumiyoshi Myouga, Tetsuro Toyoda, Wen-Hsiung Li, and Kazuo Shinozaki. Evolutionary persistence of functional compensation by duplicate genes in arabidopsis. *Genome Biology and Evolution*, 1:409–414, 2009.

12. Takashi Makino, Karsten Hokamp, and Aoife McLysaght. The complex relationship of gene duplication and essentiality. *Trends in Genetics*, 25(4):152–155, 2009.

13. Sidi Chen, Yong E Zhang, and Manyuan Long. New genes in Drosophila quickly become essential. *Science*, 330(6011):1682–1685, 2010.

14. Shane Woods, Avril Coghlan, David Rivers, Tobias Warnecke, Sean J Jeffries, Taejoon Kwon, Anthony Rogers, Laurence D Hurst, and Julie Ahringer. Duplication and retention biases of essential and non-essential genes revealed by systematic knockdown analyses. *PLoS Genetics*, 9(5):e1003330, 2013.

15. Raquel Assis and Doris Bachtrog. Neofunctionalization of young duplicate genes in Drosophila. *Proceedings of the National Academy of Sciences*, 110(43):17409–17414, 2013.

16. Raquel Assis and Doris Bachtrog. Gradual divergence and diversification of mammalian duplicate gene functions. *bioRxiv*, 2014.

17. John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (GTEx) project. *Nature Genetics*, 45(6):580–585, 2013.

18. T. Babak, B. DeVeale, E. Tsang, Y. Zhou, X. Li, K.S. Smith, K.R. Kukurba, R. Zhang, J.B. Li, D. van der Kooy, S.B. Montgomery, and H.B. Fraser. Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. *Nature Genetics*, 2015, [Epub ahead of print].

19. James R Lupski. Hotspots of homologous recombination in the human genome: not all homologous sequences are equal. *Genome Biology*, 5:2004–5, 2004.

20. Bryce van de Geijn, Graham McVicker, Yoav Gilad, and Jonathan Pritchard. WASP: allele-specific software for robust discovery of molecular quantitative trait loci. *bioRxiv*, page 011221, 2014.

21. Matthew J Lambert, Wayne O Cochran, Brandon M Wilde, Kyle G Olsen, and Cynthia D Cooper. Evidence for widespread subfunctionalization of splice forms in vertebrate genomes. *Genome research*, 2015.

22. Kai Peng, Wei Xu, Jianyong Zheng, Kegui Huang, Huisong Wang, Jiansong Tong, Zhifeng Lin, Jun Liu, Wenqing Cheng, Dong Fu, et al. The disease and gene annotations (DGA): an annotation resource for human disease. *Nucleic Acids Research*, 41:D553–D560, 2013.

23. Wenqing Fu, Timothy D OConnor, Goo Jun, Hyun Min Kang, Goncalo Abecasis, Suzanne M Leal, Stacey Gabriel, David Altshuler, Jay Shendure, Deborah A Nickerson, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 2012.

24. David Brawand, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, Angélica Liechti, Ayinuer Aximu-Petri, Martin Kircher, et al. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348, 2011.

25. Wenfeng Qian, Ben-Yang Liao, Andrew Ying-Fei Chang, and Jianzhi Zhang. Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends in Genetics*, 26(10):425–430, 2010.

26. Lluís Armengol, Miguel Angel Pujana, Joseph Cheung, Stephen W Scherer, and Xavier Estivill. Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Human molecular genetics*, 12(17):2201–2208, 2003.

27. Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

28. Avazeh Ghanbarian and Laurence Hurst. Neighboring genes show correlated evolution in gene expression. *Molecular Biology and Evolution*, 2015.

29. Mira V Han, Jeffery P Demuth, Casey L McGrath, Claudio Casola, and Matthew W Hahn. Adaptive evolution of young gene duplicates in mammals. *Genome Research*, 19(5):859–867, 2009.

30. Megan Y Dennis, Xander Nuttle, Peter H Sudmant, Francesca Antonacci, Tina A Graves, Mikhail Nefedov, Jill A Rosenfeld, Saba Sajjadian, Maika Malig, Holland Kotkiewicz, et al. Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell*, 149(4):912–922, 2012.

31. Cécile Charrier, Kaumudi Joshi, Jaeda Coutinho-Budd, Ji-Eun Kim, Nelle Lambert, Jacqueline De Marchena, Wei-Lin Jin, Pierre Vanderhaeghen, Anirvan Ghosh, Takayuki Sassa, et al. Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell*, 149(4):923–935, 2012.

32. Leonardo Salmena, Laura Poliseno, Yvonne Tay, Lev Kats, and Pier Paolo Pandolfi. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell*, 146(3):353–358, 2011.

33. Konstantin Y Popadin, Maria Gutierrez-Arcelus, Tuuli Lappalainen, Alfonso Buil, Julia Steinberg, Sergey I Nikolaev, Samuel W Lukowski, Georgii A Bazykin, Vladimir B Seplyarskiy, Panagiotis Ioannidis, et al. Gene age predicts the strength of purifying selection acting on gene expression variation in humans. *The American Journal of Human Genetics*, 95(6):660–674, 2014.

Acknowledgements. This work was funded by NIH grants ES025009 and MH101825, and by the Howard Hughes Medical Institute. We thank the GTEx Consortium and Hunter

Fraser for prepublication access to expression data, and Dmitri Petrov, Hunter Fraser, Patrick Phillips, Molly Przeworski, Arbel Harpak, Yang I. Li and Audrey Fu for comments and discussion.