Contribution of Nonohnologous Duplicated Genes to High Habitat Variability in Mammals

Satoshi C. Tamate,1 Masakado Kawata,*1 and Takashi Makino*1
1Department of Ecology and Evolutionary Biology, Graduate School of Life Sciences, Tohoku University, Aoba-ku, Sendai, Japan
*Corresponding author: E-mail: kawata@m.tohoku.ac.jp; tamakino@m.tohoku.ac.jp.

Abstract

The mechanism by which genetic systems affect environmental adaptation is a focus of considerable attention in the fields of ecology, evolution, and conservation. However, the genomic characteristics that constrain adaptive evolution have remained unknown. A recent study showed that the proportion of duplicated genes in whole Drosophila genomes correlated with environmental variability within habitat, but it remains unclear whether the correlation is observed even in vertebrates whose genomes including a large number of duplicated genes generated by whole-genome duplication (WGD). Here, we focus on fully sequenced mammalian genomes that experienced WGD in early vertebrate lineages and show that the proportion of small-scale duplication (SSD) genes in the genome, but not that of WGD genes, is significantly correlated with habitat variability. Moreover, species with low habitat variability have a higher proportion of lost duplicated genes, particularly SSD genes, than those with high habitat variability. These results indicate that species that inhabit variable environments may maintain more SSD genes in their genomes and suggest that SSD genes are important for adapting to novel environments and surviving environmental changes. These insights may be applied to predicting invasive and endangered species.

Key words: habitat diversity, evolvability, adaptation, gene duplication, whole-genome duplication.

Introduction

The areas of habitats and environmental variability within habitats considerably vary with species. Some species have large habitat ranges that include various climatic conditions and vegetation types, whereas other species have narrow ranges that include only one type of environmental condition. Although geographical and ecological factors affect the range limits, evolutionary factors affecting the capacity for adaptation to novel and different environments are important influences on range size and environmental variability within habitats (Kirkpatrick and Barton 1997). For the genetic factors, both a lack and an excess of genetic variation can restrict their expansion to novel environments within habitat boundaries (Hoffmann and Blows 1994). A recent study reported that a Drosophila population with low genetic variation had low cold and desiccation tolerance and narrow habitat distribution compared with a population with high genetic variation (Kellermann et al. 2009). These observations indicate that lack of genetic variation is related to failure in adapting to novel environments.

Makino and Kawata (2012) focused on gene duplication as the source of genome-wide genetic variation for adapting various environments. Using 11 fully sequenced Drosophila species, the authors showed that species with high proportion of duplicated genes (P_D) in a genome inhabit locations with high environmental variability and proposed that P_D in a genome is correlated with adaptation to various environments in Drosophila. Gene duplication, particularly common in eukaryotes, generates redundant gene copies, and mutations are likely to accumulate in duplicated genes under relaxed functional constraints. Although most duplicated genes are immediately lost under these relaxed constraints, functionally redundant duplicated genes can be maintained for a long time (Ohno 1970; Dean et al. 2008), resulting in the accumulation of genetic variation. Genetic variation in duplicated genes within a population is likely to be maintained by their buffering effect (Wilkins 1997; Hartman et al. 2001). Furthermore, a theoretical study showed that duplicated genes are likely to be maintained in gene regulatory networks in randomly fluctuating environments (Tsuda and Kawata 2010). Maintained duplicated genes differentiate their function and/or expression pattern. Thus, gene duplication is a major source of genetic variation in a genome. Although a positive correlation between P_D and habitat diversity was observed in Drosophila species, it is still unclear whether the tendency holds true in other species.

Some groups of eukaryotes including vertebrates, plants, and fungi have experienced whole-genome duplication (WGD), and accordingly carry two types of duplicated genes, one generated by WGD and the other by small-scale duplication (SSD). Ohnologs, which are duplicated genes generated by WGD (Wolfe 2000), are not just ancient duplicated remnants and have different features compared with SSD genes (Maere et al. 2005; Blomme et al. 2006; Hakes et al. 2007; Wapinski et al. 2007). Ohnologs are likely to be dosage-balanced genes, and thus rarely experienced gene duplications and losses during evolution (Veitia 2002, 2003, 2004; Birchler et al. 2005; Makino and McLysaght 2010; Birchler and Veitia 2012). Gene knockouts of an ohnolog causes lethal phenotypes compared with that of SSD genes as old as ohnologs (Makino et al. 2009). Therefore, ohnologs are unlikely to display not only copy number variations (CNVs) of themselves in a human population (Makino and
McLysaght et al. (2010) but also CNVs of their neighbor genomic regions (Makino et al. 2013). Ohnologs are also associated with dosage-sensitive functions, such as development, transcription regulation, and protein complex formation (Maere et al. 2005; Blomme et al. 2006; Brunet et al. 2006; Hufton et al. 2008; Makino et al. 2009). SSD genes, in contrast, are possibly related to response to biotic stimuli, a property that could be important for adaptive traits (Blomme et al. 2006). In addition, Satake et al. (2012) showed that tissue-specific expression of SSD genes was overrepresented in endoderm tissues (stomach, colon, liver, etc.) directly facing the outer environment, whereas tissue-specific ohnologs were likely to be expressed in ectoderm tissues (nerves system, brain, eye, etc.) not exposed to the outer environment. We accordingly hypothesized that SSD genes contribute more to high environmental adaptability than ohnologs.

Drosophila do not experience WGD, so the effects of the different types of duplicated genes on environmental adaptability remain unknown. Mammalian species, in contrast, experienced WGD twice in an early vertebrate lineage (McLysaght et al. 2002; Dehal and Boore 2005; Nakatani et al. 2007). The purpose of this study was to investigate whether \( P_D \) correlates with environmental variability even in mammals and to test our hypothesis that the proportion of SSD genes has a larger effect on the environmental variability than the proportion of ohnologs. Using 30 fully sequenced mammalian species, we performed comparative genomic analyses to investigate the relationship between duplicated genes—either ohnologs or SSD genes—and environmental variability.

**Results and Discussion**

**\( P_D \) Associated with Habitat Diversity**

A previous research reported a relationship between \( P_D \) and habitat diversity for closely related species in the same genus (Makino and Kawata 2012). The fully sequenced mammalian species used in our study were classified in only one genus per order except for euarchontoglires including Primates (nine species), Rodentia (five species), and Lagomorpha (two species). Thus, we used 16 species of euarchontoglires to investigate the relationship between genomic architecture and habitat variability, employing a linear model in which genome size, genome coverage and the total number of genes reported by Ensembl (Flicek et al. 2011), the number of duplicated genes, and \( P_D \) were used as explanatory variables, and the climatic envelope (see Materials and Methods) was used as a response variable. As a result of the model selection, only \( P_D \) was selected as an explanatory variable. When the effects of the climatic envelope, habitat area, and diet breadth on \( P_D \) were examined, only the climatic envelope was selected as an explanatory variable. When we investigated the relationship between \( P_D \) and the climatic envelope, there was a significant positive relationship of \( P_D \) with the climatic envelope (\( R^2 = 0.32, P = 0.012 \); fig. 1A and supplementary table S1, Supplementary Material online) although there was no correlation between \( P_D \) and the climatic envelope over all 30 mammals (supplementary fig. S1, Supplementary Material online). This result indicates that \( P_D \) correlates with the climatic envelope for closely related species (euarchontoglires in fig. 1A and Drosophila [Makino and Kawata 2012]), but the relationship was disrupted during evolution resulting in no correlation for distantly related species (mammals overall).

It has been reported that duplicated genes consist of many more recently duplicated genes than old duplicates (Lynch and Conery 2000). This disparity could be caused by high rates of gene duplications and losses (Perriere and Gouy 1996). The same trend was observed in a previous study of Drosophila (Makino and Kawata 2012) in which many duplicated gene pairs tended to have the low number of substitutions per synonymous site (\( K_S \) < 0.1. Many nondivergent duplicated

**Fig. 1.** Correlation between \( P_D \) and the climatic envelope for euarchontoglires including Primates, Rodentia, and Lagomorpha. (A) Relationship between \( P_D \) and the climatic envelope. The x-axis represents PICs in the \( P_D \) for 16 euarchontoglires. The y-axis represents PICs in the climatic envelope estimated from WORLDCLIM data sets. (B) Relationship between proportions of lost duplicated genes and the climatic envelope. The x-axis represents PICs in the proportion of lost duplicated genes for 16 euarchontoglire species. The y-axis represents PICs in the climatic envelope estimated from WORLDCLIM data sets.
gene pairs would be found as a result of recent duplication events or assembly errors in a genome. Therefore, we collapsed recent duplicated genes based on their sequence similarity. We estimated \( K_p \) between duplicated genes and their closest paralogs in mammals and found an apparent distribution bias toward recent duplication (supplementary fig. S2, Supplementary Material online). We accordingly collapsed duplicated gene pairs with a \( K_p < 0.1 \) into single genes (see Materials and Methods). Although there was no correlation between the proportion of recently duplicated genes and the climatic envelope (supplementary fig. S3A, Supplementary Material online), there was a significant correlation between \( P_D \) and the climatic envelope after collapsing recently duplicated genes (\( R^2 = 0.31, P = 0.013 \); supplementary fig. S3B, Supplementary Material online). These results indicate that recently duplicated genes do not explain the habitat variability of species because their duplication is too recent to have generated genetic variation, otherwise assembly errors might disturb the relationship. Note that even when we set some threshold values of \( K_p (0.01–0.10) \) for collapsing duplicated genes, our result did not change (data not shown).

Some mammalian species’ habitats have been expanded by human activity. Among species used in this study, \( M. \) *musculus* in particular has spread worldwide. When the same analysis was conducted after the removal of \( M. \) *musculus*, the trends did not change (the correlation between \( P_D \) and the climatic envelope: \( R^2 = 0.35, P = 0.011 \); the correlation between \( P_D \) and the climatic envelope after collapsing recently duplicated genes: \( R^2 = 0.35, P = 0.012 \); supplementary fig. S4, Supplementary Material online). We did not use \( M. \) *musculus* in the subsequent analyses, although our conclusions were unaffected by the inclusion of this species.

Some other factors might have affected our results. Although a correlation between effective population size and genomic content has been reported (Petit and Barbadilla 2009), the difference in \( P_D \) among species was not related to effective population sizes in *Drosophila* species (Makino and Kawata 2012). For the mammalian species used in this study, we could not estimate effective population size; this correlation should be evaluated in future studies. Differences in genome sequence coverage among species might have influenced our results; however, there was no correlation between genome sequence coverage and \( P_D \) (supplementary fig. S5, Supplementary Material online) as shown in a previous study (Makino and Kawata 2012).

To reinforce the above results, the habitat diversity defined by Köppen–Geiger climate classification was used instead of the climatic envelope. This classification is based not only on temperature and precipitation but also on vegetation (Huijser 1999). Environmental diversity within habitat was estimated using the Brillouin index (Pielou 1975; Kottek et al. 2006). When this habitat diversity was used as a measure of environment variability, a similar result was obtained (\( R^2 = 0.23, P = 0.034 \); supplementary fig. S6A, Supplementary Material online). The result did not change when \( P_D \) after collapsing recently duplicated genes was used (\( R^2 = 0.22, P = 0.037 \), supplementary fig. S6B, Supplementary Material online). This result indicates that \( P_D \) is correlated with habitat variability not only in *Drosophila* species but also in euarchontoglires.

### Evolutionary Process on Divergence of Duplicated Genes

We have already shown that recently duplicated genes were not enriched in species with high habitat variability (supplementary fig. S3A, Supplementary Material online). To identify evolutionary processes responsible for differences in \( P_D \) between species with high and low habitat variability, we investigated whether in euarchontoglires species a loss of duplicated genes occurred in species with low \( P_D \) more frequently than in those with high \( P_D \). Among these species, there was a significant negative correlation between the proportion of lost duplicated genes and the climatic envelope (\( R^2 = 0.51, P = 0.0030 \); fig. 1B). The result is consistent with a previous report on *Drosophila* (Makino and Kawata 2012) in which species with low habitat variability tended to lose duplicated genes during evolution.

Overrepresented gene functional categories were examined for lost genes by Gene Ontology (GO) analysis (see Materials and Methods) with two closely related species pairs (*Pongo abelii–Macaca mulatta* and *Spermophilus tridecemlineatus–M. musculus*) in euarchontoglires. *Pongo abelii* and *S. tridecemlineatus* are species with low \( P_D \) and habitat variability, whereas *M. mulatta* and *M. musculus* are species with high \( P_D \) and habitat variability (table 1). The results showed little enrichment of functional categories for genes in species with low \( P_D \); the lost duplicated genes in *P. abelii* were enriched in catabolic process (\( P = 3.6 \times 10^{-17} \)). There were no enriched gene functions in *S. tridecemlineatus*. This finding indicates that both loss and relaxation of functional

<table>
<thead>
<tr>
<th>Species Outgroup</th>
<th>Habitat Variability</th>
<th>Proportion of Duplicated Genes (( P_D ))</th>
<th>Number of Lineage-Specific Lost Genes</th>
<th>Proportion of Lineage-Specific Duplicated Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Macaca mulatta</strong> Callithrix jacchus</td>
<td>590</td>
<td>2.238</td>
<td>0.81</td>
<td>1,312</td>
</tr>
<tr>
<td><strong>Pongo abelii</strong></td>
<td>45</td>
<td>0</td>
<td>0.77</td>
<td>1,412</td>
</tr>
<tr>
<td><strong>P value</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mus musculus</strong> Ochotonan princeps</td>
<td>1,201</td>
<td>3.067</td>
<td>0.81</td>
<td>584</td>
</tr>
<tr>
<td><strong>Spermophilus tridecemlineatus</strong></td>
<td>155</td>
<td>1.718</td>
<td>0.68</td>
<td>401</td>
</tr>
</tbody>
</table>

**Table 1.** Lineage-Specific Evolutionary Events between Closely Related Species.
constraints are common for genes in species with low habitat variability, rather than being gene-specific. For adapting to heterogeneous environments, species might require not only some particular biological functions such as cold and desiccation tolerance but also physiological, morphological, behavioral, and other adaptive functions. The present results support the earlier conclusion (Makino and Kawata 2012) that duplicated genes have been lost from the genome in species with low habitat variability, whereas high \( P_D \) has been maintained in the genome in species with high habitat variability.

**Influence of Different Duplication Mechanisms**

When the effects of ohnologs and SSD genes on habitat variability were separately examined, the proportion of SSD genes significantly correlated with habitat variability \((R^2 = 0.39, P = 0.013; \text{fig. 2A})\), whereas the proportion of ohnologs did not \((\text{fig. 2B})\). The results suggest that the proportion of SSD genes makes a larger contribution to adaptation to variable environments than the proportion of ohnologs although \( P_D \) calculated using whole genes could be a factor that restricts adaptation to environmental variation as shown in *Drosophila* species (Makino and Kawata 2012). Ohnologs would be unlikely to be lost under dosage sensitivity regardless of species' habitat variability (Makino and McLysaght 2010). In contrast, species with high habitat variability could maintain SSD genes more efficiently, whereas species with low habitat variability would lose SSD genes more frequently. As a result, species that had lost SSD genes might be unable to generate enough genetic variation for adaptation to heterogeneous environments, leading to failure to expand their habitat range, resulting in further losses of SSD genes. Thus, SSD genes can be both a cause and effect of a change in habitat variability. SSD genes are related to response to biotic stimuli for adaptive traits and were overrepresented in endoderm tissues directly exposed to external environments (Blomme et al. 2006; Satake et al. 2012). Thus, retained SSD genes that have responded to various environmental stimuli have played more significant roles in expanding habitat ranges and in adaptation to novel environments than ohnologs.

**Pairwise Comparison for Mammals**

In this study, no correlation between \( P_D \) and habitat variability was observed when 30 mammalian species were used. But a positive correlation might be observed when species within the same taxonomic group (e.g., species in the same order) were compared, as in *Drosophila* (Makino and Kawata 2012) and euarchontoglires. To investigate the relationship between \( P_D \) and habitat variability for other mammalian taxa, pairs of species were classified by shared membership in an order, a related order, or a superorder \((\text{supplementary table S2, Supplementary Material online})\). Figure 3A depicts the relationship between \( P_D \) and the climatic envelope in each species pair. We found that within pairs, most species with high \( P_D \) had higher climatic envelopes than species with low \( P_D \) \((\text{exact binomial test,} P = 0.011)\). In other words, most closely related species pairs of mammalian species showed a positive correlation between \( P_D \) and habitat variability. Figure 3B depicts a similar analysis of SSD genes and the climatic envelope, showing a significant correlation \((\text{exact binomial test,} P = 0.013; \text{fig. 3B})\). In contrast, there was no significant correlation between the proportion of ohnologs and the climatic envelope \((\text{exact binomial test,} P = 0.18; \text{fig. 3C})\). These results suggest that \( P_D \), particularly of SSD genes, could be a general index of species' capacity to adapt to habitat variability. *Rattus norvegicus* has high habitat variability \((\text{fig. 3})\), and it might affect to the above results. Although the same analyses were

![Fig. 2](image_url)
conducted after the removal of *R. norvegicus*, the trends did not change (exact binomial test, $P = 0.039$ in fig. 3A, $P = 0.048$ in fig. 3B and $P = 0.55$ in fig. 3C). Exceptionally, there were some species that showed negative relationships between $P_D$ and the climatic envelope (*Myotis lucifugus–Pteropus vampyrus* and *Echinops telfairi–Procavia capensis*; fig. 3A). Estimation of environmental variability would reflect only broad-scale patterns. In addition, the measure of habitat ranges of species is not perfect.

**Prospective Index for Evaluating Evolvability**

The present results using fully sequenced genomes of mammalian species support our hypothesis that $P_D$ is correlated with environmental variability within the species’ habitat, as previously shown for *Drosophila*. Furthermore, our results showed that it is mainly the proportion of SSD genes (rather than that of ohnologs) that contributes to this correlation. Our results also showed that species with low habitat variability and low $P_D$ tended to lose duplicated genes.

![Graphs](image-url)
We suggest that the proportion of SSD genes may be both a cause and an effect of habitat variability in mammalian species, although further studies would be needed to explore how maintained SSD genes generate a capacity for adaptation to various environments. Whole genomes can now be sequenced relatively easily and the sequences have been accumulated in databases, so that in the near future we will be able to estimate $P_D$ for thousands of species (Genome 10 K, https://genome10k.soee.uchsc.edu, last accessed April 14, 2014 and i5k Insect http://www.arthropodgenomes.org/wiki/i5K, last accessed April 14, 2014). $P_D$ in particular the proportion of SSD genes, would be an appropriate index for evaluating species vulnerability to future environmental changes for the conservation of biodiversity.

**Materials and Methods**

**Sequences of Fully Sequenced Mammalian Species**

following were observed: an inferred ortholog (species2A in supplementary fig. S7B, Supplementary Material online) in the compared species (species2 in supplementary fig. S7B, Supplementary Material online) was a duplicated gene and the similarity between the duplicated gene and its duplicated gene partner (similarity between species2A and species2B in supplementary fig. S7B, Supplementary Material online) was determined by BLAST search.

Habitat Area and Variability
The habitat areas for mammalian species were obtained from the International Union for Conservation of Natural Resources (IUCN) Red List of Threatened Species and literatures (http://www.iucnredlist.org, last accessed April 14, 2014) (Huijser 1999; Marroig et al. 2004). Habitat variability was estimated from the climatic envelope, and habitat diversity was estimated using the Köppen climatic classification (Kottek et al. 2006). The climatic envelope is the range of temperatures, precipitation, and other climate-related parameters in which a species is currently found (supplementary table S3, Supplementary Material online). We estimated the climatic envelope using principal component analysis (PCA) with WORLDCLIM (Hijmans et al. 2005). We obtained world special data and the WORLDCLIM climatic data set (10 min latitude/longitude) from DIVA-GIS (http://www.diva-gis.org, last accessed April 14, 2014). The habitat area was measured as the number of grid squares on the climate map. We then extracted the climatic values from 19 bioclimatic variables used for BIOCLIM (Hijmans et al. 2005) in the habitat area of each mammalian species. We performed PCA using the bioclimatic variables for all of the species and found that the first two principal components (PCs) explained 93.6% of the total variance (supplementary table S3, Supplementary Material online). The contributions of PC1 and PC2 were 77.9% and 15.7%, respectively. PCA plots (x-axis: PC1; y-axis: PC2) are shown in supplementary figure S58, Supplementary Material online. On the basis of the PCA results, PC1 and PC2 values were plotted for each species. A total of 122,303 cells were used (PC1: 779×PC2: 157) by weighting the relative contributions to PC1 and PC2 for estimating the climatic envelope, and the numbers of cell grid overlapping points in the 122,303 cell grids were defined as the climatic envelope of mammalian species.

To remove any phylogenetic constraints on the relationship between genetic architecture and habitat, phylogenetic trees and phylogenetic independent contrasts (PICs) were inferred. Phylogenetic trees were inferred using matrix extracellular phosphoglycoprotein precursor gene (Bardet et al. 2010). The ClustalW2 multiple sequence alignment program (Larkin et al. 2007) was used for alignment and NJplot was used for constructing phylogenetic trees (Perriere and Gouy 1996). Using the phylogenetic trees, we selected a model by applying a generalized least squares model with a Brownian model, as described earlier, and measured the PICs (Felsenstein 1985). The model selection was applied using the estimated PICs.

To evaluate the relationship between PD and habitat variability in species paired by common order, common superorder, and related order, we performed pairwise comparisons using a sign test.

Gene Ontology
To determine whether lineage-specific lost duplicated genes of species with low PD were enriched in specific functional categories, we examined the GO database entries for the duplicated genes between species with different PD (C. jacchus–P. abelii and M. musculus–S. tridecemlineatus) (table 1). The GO identifiers (IDs) and GO “slim” biological process annotations for M. musculus and Homo sapiens were downloaded from ftp://ftp.geneontology.org/pub/go/gene-associations/ (last accessed April 14, 2014) and http://www.geneontology.org/GO_slims (last accessed April 14, 2014), respectively. The frequency of each GO ID was counted for genes from M. musculus and H. sapiens. For the other species, we used the GO IDs of the most similar homolog in M. musculus (for Rodentia and Lagomorpha) and H. sapiens (for Primates). The enrichment of GO IDs for genes in species with low PD was compared with that in species with high PD. We calculated the P value for each GO ID by using Brillouin’s index, which is robust to sample size (Margalef 1958).

Model Selection
All of the following statistical analyses were executed with R software (2.13.2 version) (http://www.r-project.org, last accessed April 14, 2014). Model selection was applied using linear models to determine which genomic factors affect habitat features. The set of predictors of the explanatory variables that yielded the lowest Akaike Information Criterion (AIC) was selected using a stepwise AIC procedure. PD, PD for SSD genes, PD for orthologs, genome size, the number of genes, and the number of duplicated genes as genomic factors were used, and the climatic envelope, habitat diversity, diet breadth, and habitat area were used as habitat features. Diet breadth is the number of dietary categories eaten by each species measured using any qualitative or quantitative dietary measure over any period of time using any assessment method for noncaptive or nonprovisioned populations (Jones et al. 2009).
comparing the two different gene sets. The estimated P values were adjusted by Bonferroni correction.

**Supplementary Material**

Supplementary tables S1–S3 and figures S1–S8 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

The authors thank our laboratory members for their helpful advice. This work was supported by KAKENHI (25291096) from the Japan Society for the Promotion of Science to T.M., and partly by the Global COE Program “Centre for ecosystem management adapting to global change” (J03) of the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) to M.K.

**References**


Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. 2007. All duplications are not created equal: the difference between small-scale and genome duplication. *Genome Biol.* 8:R209.


