Gene Duplication in the Honeybee: Patterns of DNA Methylation, Gene Expression, and Genomic Environment

Carl J. Dyson[1] and Michael A.D. Goodisman *,[1]

[1]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA

***Corresponding author:** E-mail: michael.goodisman@biology.gatech.edu.

Associate editor: Naruya Saitou

Abstract

Gene duplication serves a critical role in evolutionary adaptation by providing genetic raw material to the genome. The evolution of duplicated genes may be influenced by epigenetic processes such as DNA methylation, which affects gene function in some taxa. However, the manner in which DNA methylation affects duplicated genes is not well understood. We studied duplicated genes in the honeybee *Apis mellifera*, an insect with a highly sophisticated social structure, to investigate whether DNA methylation was associated with gene duplication and genic evolution. We found that levels of gene body methylation were significantly lower in duplicate genes than in single-copy genes, implicating a possible role of DNA methylation in postduplication gene maintenance. Additionally, we discovered associations of gene body methylation with the location, length, and time since divergence of paralogous genes. We also found that divergence in DNA methylation was associated with divergence in gene expression in paralogs, although the relationship was not completely consistent with a direct link between DNA methylation and gene expression. Overall, our results provide further insight into genic methylation and how its association with duplicate genes might facilitate evolutionary processes and adaptation.

***Key words:* DNA methylation, epigenetics, Hymenoptera, *Apis mellifera*, gene expression, molecular evolution.**

Introduction

Gene duplication is a fundamentally important process that introduces new genetic material into the genome (Conant and Wolfe 2008). New genes generated through duplication events can be expressed in novel ways (Holland et al. 1994; Li et al. 2018). For example, after duplication, one paralog (duplicate gene copy) can be removed from the selective pressure of its ancestral function and be expressed in a divergent fashion from its sister paralog (Stephens 1951; Nei 1969; Ohno 1970; Otto and Whitton 2000).

Paralogous genes can evolve in several ways following duplication events. The majority of duplicate gene copies become nonfunctionalized and are removed through negative selection (Lynch and Conery 2000). However, in some circumstances, both paralogs persist. In such cases, one possible outcome is the conservation of ancestral function in both duplicated genes, allowing for amplification of that function through gene dosage effects. Alternatively, a duplicated gene can serve a novel function by gaining a new expression profile. Finally, paralogs may experience subfunctionalization, which requires expression of both copies to fulfill the original ancestral function and level of expression (Ohno 1970; Force et al. 1999). The distinct fates of duplicate genes provide the mechanisms for paralogs to display distinct expression profiles, allowing for the possibility of adaptation and evolution of organismal function.

One mechanism by which the expression of duplicate genes could be regulated is through the effects of DNA methylation (Berger et al. 2009). DNA methylation is an heritable epigenetic modification that is found across a wide array of species including animals, plants, bacteria, and fungi (Suzuki and Bird 2008; Feng et al. 2010; Niederhuth et al. 2016; Bewick, Zhang, et al. 2019). Associations between DNA methylation and gene expression have been uncovered previously. For example, methylation of certain targeted genomic regulatory elements, namely the promoter regions of genes, has been causatively associated with a repression of gene expression levels in vertebrate systems (Jones 2012).

The function of methylation of gene bodies, which is the predominant target of DNA methylation in insects, is less clear (Zilberman 2017). Gene body methylation has been proposed to affect gene expression through regulation of transcription elongation and alternative splicing (Bird 2002; Lorincz et al. 2004; Zilberman and Henikoff 2007; Luco et al. 2010; Maunakea et al. 2010; Shukla et al. 2011). DNA methylation in many insect species is found at much lower overall levels than in vertebrate and plant species and is mainly localized to cytosine-guanine (CpG) dinucleotides in intragenic regions, rather than being found across the genome (Wang et al. 2006; Zemach et al. 2010). Additionally, the sparse DNA methylation in holometabolous insects typically targets phylogenetically conserved "housekeeping" genes (Elango et al. 2009; Foret et al. 2009; Lyko et al. 2010; Sarda et al. 2012). These methylated genes are associated with constitutive expression across tissues and phenotypes (Hunt et al. 2010; Bonasio et al. 2012; Glastad et al. 2013). Thus, gene body methylation is generally positively associated with levels of gene expression and negatively associated with gene expression bias. However, the causal connection, if any, between gene expression and gene body methylation remains unclear (Simmen et al. 1999; Suzuki and Bird 2008; Feng et al. 2010; Zemach et al. 2010; Bewick, Sanchez, et al. 2019; Bewick, Zhang, et al. 2019).

The potential effects of DNA methylation on the evolution of duplicate gene function have been investigated previously (Rodin and Riggs 2003; Qian et al. 2010; Ramirez-Gonzalez et al. 2018). These prior studies focused on methylation of promoter regions, which has been linked to a downregulation in gene expression in vertebrate systems. Results from these investigations suggest that it would be possible for epigenetic modification to maintain paralog expression after a duplication event, allowing for subsequent functional modification and differential expression. Additionally, previous work indicates promoter DNA methylation as a causative factor in tissue-specific expression of duplicate genes (Keller and Yi 2014). This suggests that DNA methylation could play a role in the evolution of duplicate genes in eukaryotic species.

The study of gene duplication and epigenetics holds particular interest in social insects (Kucharski et al. 2016). Social insects are remarkable because they display a caste system, where members of the same species share a common genotype yet express strikingly different phenotypes (Wheeler 1986; Normark 2003; Simpson et al. 2011). Social insects represent interesting model systems, because gene duplication may have important effects on the elaboration of castes. For example, selection should favor each caste member to evolve a phenotype most beneficial to its own fitness requirements. However, if castes have different selective requirements for expression at a single locus, genetic conflicts can emerge (Bonduriansky and Chenoweth 2009; Connallon and Clark 2011; Pennell et al. 2018). This conflict can be mitigated through gene duplication (Ellegren and Parsch 2007; Gallach and Betran 2011). The insertion of novel genetic material into the genome through duplication allows for the evolution of different expression profiles for a single gene, providing the means to reach the fitness maximum for each class.

The purpose of this study is to gain a greater understanding of the consequences of gene duplication by identifying how evolutionary and epigenetic processes affect duplicate genes. We study this question in the honeybee *Apis mellifera*, which displays a defined caste system where the expression of genes in multiple copy could hold significance to social evolution. We are specifically interested in understanding if DNA methylation is associated with gene duplication. To examine this issue, we investigate the association between levels of genic methylation and gene duplication status in a variety of genic contexts. Additionally, we investigate divergence in paralogs—from each other and from orthologous gene copies—to provide a framework for understanding the evolution of duplicated genes. Our research provides further insight into questions regarding the nature of DNA methylation in invertebrate systems and the role of gene duplication in a social insect.

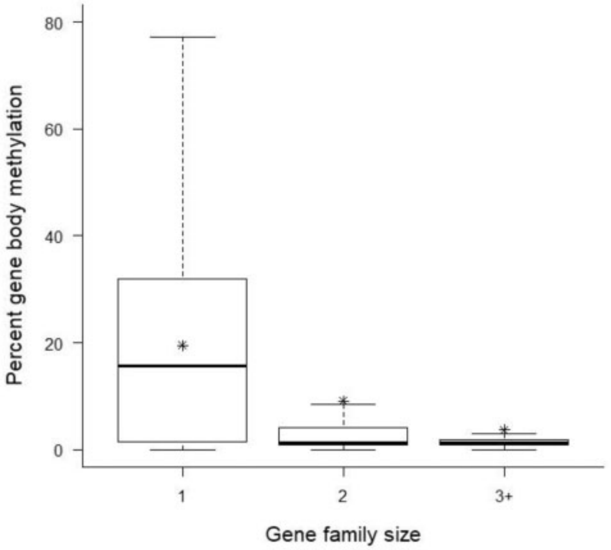

Fig. 1. Percentage of methylated CpG dinucleotides in gene families of size 1 (singletons), 2 (duplicates), and 3 or more. Box represents first quartile, median, and third quartile values, whereas whiskers represent values within 1.5× the interquartile range. Stars represent mean gene body methylation percentage for each group. All three means differed significantly ($P < 0.0001$).

Results

DNA Methylation Is Associated with Caste-Biased Gene Expression and Gene Copy Number

We compared the methylation levels of 5,235 single-copy genes, 734 duplicated genes, and 271 genes in families of three or larger in the honeybee, *A. mellifera*. We found that the mean percentage of gene body methylation differed significantly among gene family classes. Singletons were significantly more highly methylated than duplicates, which were significantly more highly methylated than genes in larger families ($F(2, 6{,}190) = 177.29$, $P < 0.0001$, fig. 1 and supplementary table S1, Supplementary Material online).

We also investigated associations between gene duplication and DNA methylation using χ^2 tests of independence. We found that the methylation status and the duplication status of paralogs were significantly associated ($\chi^2(1, N = 5{,}936) = 396.37$, $P < 0.001$, supplementary tables S2 and S3, Supplementary Material online). Specifically, there were substantially more duplicated genes showing low levels of methylation, and fewer duplicate genes showing high levels of methylation, than expected. This result further indicates that patterns of DNA methylation depend on whether a gene is a duplicate or a singleton.

We determined whether relative divergence (D_r) in methylation between paralogs differed from D_r in methylation between randomly paired "pseudoparalogs." We found that randomly paired duplicated genes displayed a mean D_r of 0.568 and a standard deviation of 0.015. Similarly, the mean methylation D_r of randomly paired singletons was 0.549 with a standard deviation of 0.006. In contrast, the mean methylation D_r between actual duplicated genes was 0.242, which fell significantly below both other distributions of D_r

Table 1. χ^2 Tests of Independence between DNA Methylation Level and Caste-Biased Gene Expression for Singletons and Duplicate Genes in Six Gene Expression Data Sets.

Data Set	Singleton		Duplicate	
	χ^2	*P* Value	χ^2	*P* Value
Drone–Queen[a]	309.65	***	2.611	NS
Drone–Worker[a]	413.97	***	15.34	***
Queen–Worker[a]	399.25	***	0.5792	NS
Drone–Queen[b]	304.35	***	0.0084	NS
Drone–Worker[b]	27.85	***	1.660	NS
Queen–Worker[b]	347.96	***	0.078	NS

Note.—NS, not significant.

[a] Ashby et al. (2016).

[b] Vleurinck et al. (2016).

***$P < 0.001$.

(supplementary fig. S1, Supplementary Material online, $P < 0.0001$). This indicates that true paralogs are significantly less diverged in methylation than would be expected for two randomly selected duplicates or singletons.

Finally, we considered associations between gene duplication status, methylation level, and gene expression bias between castes in *A. mellifera*. We found a strong association between caste-biased expression and gene body methylation in *A. mellifera* singletons. Specifically, we found an excess of genes with low methylation levels showing caste-biased expression, and an excess of highly methylated genes showing unbiased expression between castes. However, in duplicated genes, this trend was absent in five of the six analyses (table 1 and supplementary table S4, Supplementary Material online) indicating that the association of gene body methylation with gene expression depends on whether the focal gene is a singleton or a duplicate. Importantly, the observed differences in the significance of trends for duplicate and singleton genes were still present even when controlling for sample size differences between singleton and duplicate genes. We further examined the differences in expression bias between singleton and duplicate genes by measuring associations in separately analyzed gene sets consisting of low- and high-DNA methylation genes. We found that high-methylation genes generally showed a stronger association between copy number and caste-biased expression. In contrast, the association was much weaker and usually nonsignificant in genes classified as low methylation (supplementary table S5, Supplementary Material online). These results, once again, demonstrate differences in the associations between gene expression bias and gene duplication status for genes showing differences in DNA methylation levels.

Divergence in Gene Expression and DNA Methylation of Paralogs Demonstrates Evolution of Gene Function

We next investigated the patterns of DNA methylation and gene expression for pairs of paralogous genes in *A. mellifera* and their corresponding outgroup orthologs in a related bee species, *Ceratina calcarata*. We specifically identified 92 genes that were in single copy in *C. calcarata* but which were duplicated in *A. mellifera*. We then compared patterns of methylation and gene expression between the orthologous *C. calcarata* copy (*CcalA*) and the paralogous *A. mellifera* duplicates (*AmelA1* and *AmelA2*) in order to gain a greater understanding of how gene methylation and expression evolved over time.

We found that methylation levels were significantly correlated for all three gene copies (*CcalA*–*AmelA1*, Spearman $\rho = 0.6390$, $P < 0.0001$; *CcalA*–*AmelA2*, $\rho = 0.5291$, $P < 0.0001$; *AmelA2*–*AmelA1*, $\rho = 0.4995$, $P < 0.0001$). We also found that expression levels were significantly correlated for the *C. calcarata* ortholog and each of the *A. mellifera* paralogs and marginally correlated for the two *A. mellifera* paralogs themselves (*CcalA*–*AmelA1*, $\rho = 0.2624$, $P = 0.0115$; *CcalA*–*AmelA2*, $\rho = 0.4061$, $P < 0.0001$; *AmelA2*–*AmelA1*, $\rho = 0.1750$, $P = 0.0953$). These results indicated that both levels of expression and levels of methylation were relatively conserved across species.

Additionally, we found that levels of methylation and levels of gene expression were correlated for each of the individual *A. mellifera* gene copies (*AmelA1*, $\rho = 0.3142$, $P = 0.0023$; *AmelA2*, $\rho = 0.3687$, $P = 0.0003$) but not for the *CcalA* ortholog ($\rho = 0.0564$, $P = 0.595$). Notably, however, a significant correlation between methylation and expression in *CcalA* did emerge when the sample size was increased to include all genes, reinforcing associations found in previous studies ($N = 11{,}535$, $\rho = 0.1059$, $P < 0.0001$, Rehan et al. 2016). Consequently, overall, we found evidence for a correlation between gene expression and DNA methylation among gene copies.

We next investigated D_r for methylation and expression in paralog and ortholog comparisons. We found that there was no significant correlation between D_r in methylation and expression between either the paralog or the ortholog (*AmelA1*–*CcalA*, $\rho = 0.0201$, $P = 0.8498$; *AmelA2*–*CcalA*, $\rho = 0.0490$, $P = 0.6448$). This indicates that divergence in methylation between the ortholog and paralog was not predictive of divergence in expression between the ortholog and the paralog.

We next considered if D_r in methylation for the *A. mellifera* paralogs was correlated with D_r in expression of the *A. mellifera* paralogs. We found that these measures were significantly correlated ($\rho = 0.3638$, $P = 0.0004$) indicating that methylation differences in *A. mellifera* paralogs were positively related to expression differences. We then compared divergence in methylation and expression between species by examining the relationship between D_r in methylation and D_r in expression of the *C. calcarata* ortholog and the mean of the *A. mellifera* paralogs (*CcalA* – [*AmelA1* + *AmelA2*]/2)/(*CcalA* + [*AmelA1* + *AmelA2*]/2). We found that D_r of methylation between the ortholog and the mean of the two paralogs was marginally *negatively* correlated with D_r of expression between the ortholog and the mean of the two paralogs ($\rho = -0.2245$, $P = 0.0324$). Interestingly, however, we found that there was no correlation between the species level D_r estimates and the D_r of the paralogs themselves for either methylation ($\rho = 0.1846$, $P = 0.0781$) or gene expression ($\rho = -0.1486$, $P = 0.1575$), indicating that the intraparalog divergence in gene expression or DNA methylation was not associated with the *A. mellifera*–*C. calcarata* divergence.

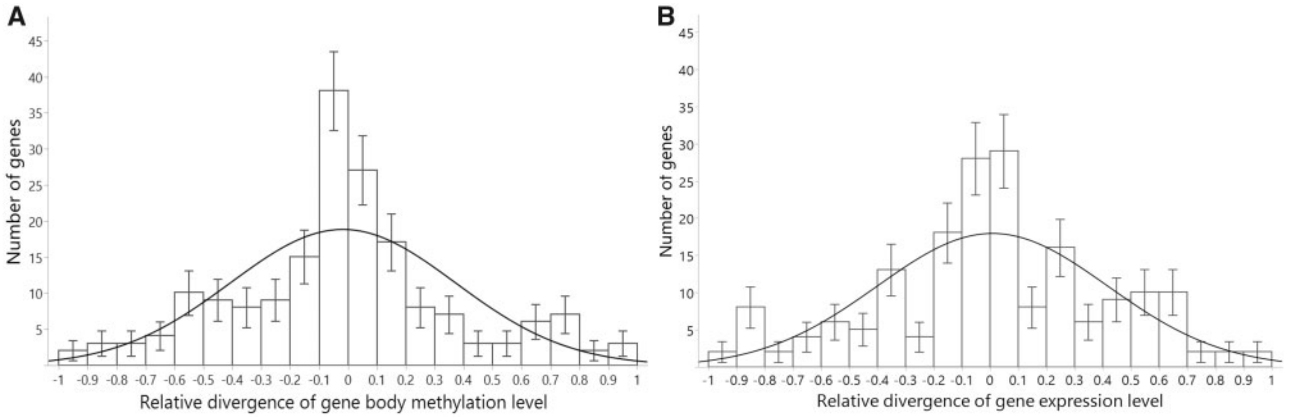

FIG. 2. Distribution of relative divergence (D_r) values for (*A*) gene methylation and (*B*) gene expression between *Apis mellifera* duplicates and their *Ceratina calcarata* singleton orthologs. Overlaid curves estimate normal distribution, and error bars represent standard error of each bin of D_r values. Both distributions of D_r values deviated significantly from Gaussian (methylation, $P = 0.0006$; expression, $P = 0.0137$).

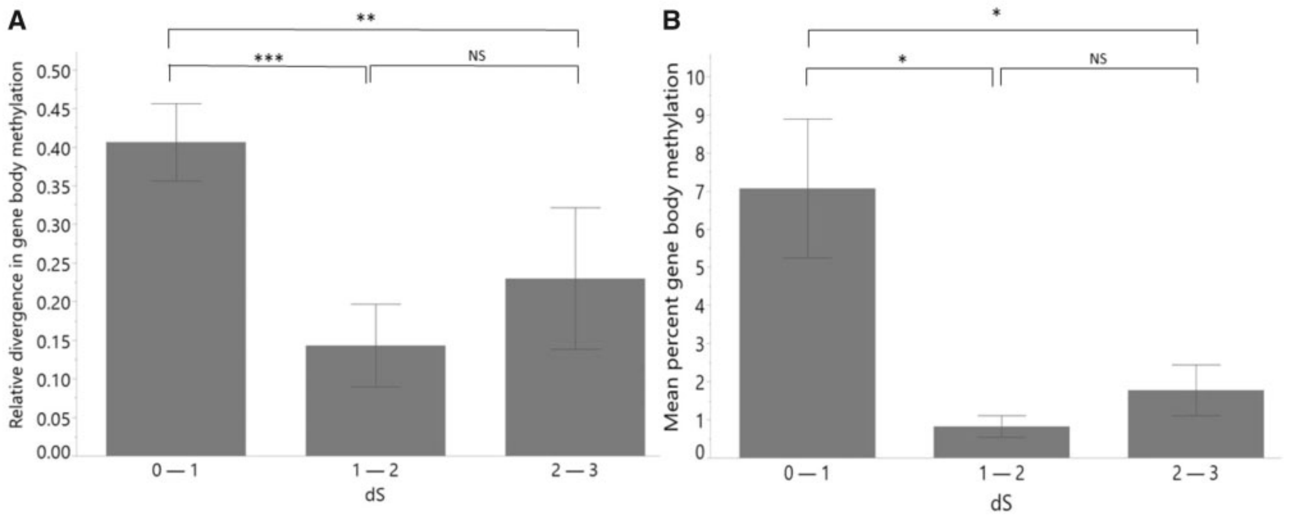

FIG. 3. (*A*) Relative divergence in DNA methylation levels for duplicated genes displaying different synonymous substitution rate ratios, indicating that younger duplicate gene pairs have more dissimilar levels of methylation than older pairs. (*B*) Mean percent gene body methylation for duplicate genes displaying different synonymous substitution rate ratios, indicating that pairs of genes that duplicated more recently tend to have higher overall methylation levels than pairs of genes that duplicated longer ago. Bars represent mean value for each bin, and error bars represent 95% confidence interval. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, and NS, not significant.

Finally, we examined the distribution of D_r values for the *C. calcarata* ortholog and the *A. mellifera* paralogs. We found that the distribution of D_r estimates for both DNA methylation and gene expression differed significantly from Gaussian (DNA methylation, $\mu = -0.0177$, $\sigma = 0.3896$, $W = 0.9704$, $P = 0.0006$, fig. 2*A;* Gene expression, $\mu = 0.0085$, $\sigma = 0.4088$, $W = 0.9811$, $P = 0.0137$, fig. 2*B*). This result suggested a nonrandom distribution of relative divergence estimates for the paralogous and orthologous gene copies.

DNA Methylation Varies with Location, Length, and Age of Duplicate Genes

We next investigated DNA sequence evolution in the context of gene duplication. Analysis of variance (ANOVA) was used to determine the strength of association between sequence divergence, as measured by dS, and the methylation divergence of all paralogs. We found that paralogs with fewer synonymous substitutions per synonymous site, which were putatively younger duplicate pairs, tended to be *more* divergent in their methylation than older duplicate pairs ($F(2, 250) = 19.1613$, $P < 0.0001$, fig. 3*A*). We also found a negative correlation between dS and the overall mean percent methylation of paralogs ($F(2, 280) = 5.1602$, $P = 0.0063$, fig. 3*B*). Together, these results indicated that older pairs of duplicate genes tended to have lower and more similar levels of DNA methylation than younger pairs.

We next tested if the methylation level of genes differed across the 16 chromosomes of the *A. mellifera* genome. We found that genes on 14 of the 16 chromosomes did not differ significantly in their mean methylation. Of the remaining two chromosomes, genes from one chromosome (chromosome 3) did not differ significantly from the means of genes on the

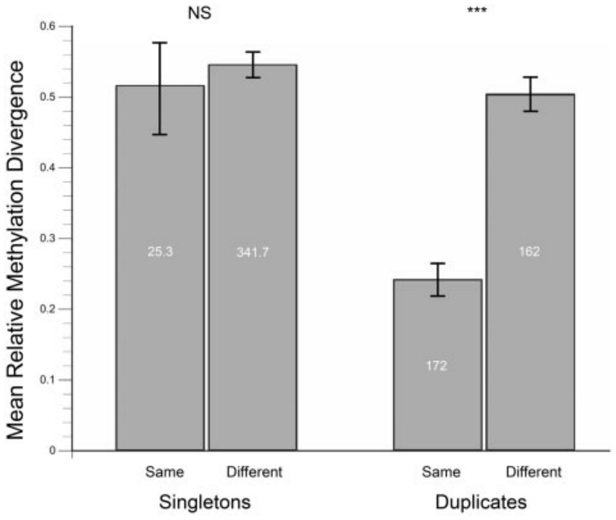

FIG. 4. Mean relative divergence in methylation for singleton pseudoparalogs (singletons) and true duplicate gene pairs (duplicates) on the same and different chromosomes. Interior-bar numbers represent number of genes in that category, and error bars represent standard error of the mean (mean of ten trials for singletons). $^{***}P < 0.0001$; NS, not significant.

13 other chromosomes. Only genes on chromosome 13 were found to differ significantly from the means of the majority of other chromosomes (supplementary fig. S2, Supplementary Material online).

We examined whether paralogous genes located on the same chromosomes (syntenic) tended to show lower levels of methylation divergence than paralogous genes located on different chromosomes (nonsyntenic). We found that nonsyntenic paralogs showed a much higher D_r in their methylation level than syntenic paralogs ($F(1, 332) = 85.9705$, $P < 0.0001$, supplementary fig. S3, Supplementary Material online). To further examine whether this association was unique to duplicated genes, we compared the divergence of randomly paired singleton “pseudoparalogs” that were syntenic or nonsyntenic. We found that D_r of methylation of these pseudoparalogous singletons did not significantly depend on whether they were found on the same or different chromosomes ($F(1, 365) = 0.6990$, $P = 0.5480$, fig. 4). Thus, the relatively low divergence in methylation of syntenic duplicate genes was not due solely to their residing on the same chromosome.

We also investigated the frequency with which duplicates resided on the same chromosome. We found that randomly paired singletons tended to fall on the same chromosome with a probability of ~1 in 16, as expected considering the 16 linkage groups to which they could belong. True duplicates, however, tended to be collocated on the same chromosome almost 50% of the time (fig. 4). A more granular inspection of the influence of chromosomal location on methylation showed that syntenic and nonsyntenic singleton pseudoparalogs had roughly the same D_r. True duplicate pairs, on the other hand, showed drastically lower mean D_r levels for pairs on the same chromosome (fig. 4).

Finally, we examined whether gene body methylation varied as a factor of gene length in both singletons and duplicated genes. We found no significant difference in the lengths of singletons and duplicated genes ($F(1, 5967) = 0.5092$, $P = 0.4755$). When examining genes up to 40 kb, applicable to over 95% of *A. mellifera* genes examined in this study, we confirmed previously uncovered negative correlations between gene body methylation and gene length in both singletons ($F(3, 5037) = 227.3255$, $P < 0.0001$) and duplicate genes ($F(3, 662) = 5.6724$, $P = 0.0008$). Interestingly, however, ANOVA of all genes binned by length in bins of 40 kb showed that the mean gene body methylation of shorter singleton genes was significantly higher than that of longer singletons ($F(7, 5{,}225) = 27.1343$, $P < 0.0001$), but this relationship was not significant in duplicate binned genes ($F(7, 694) = 1.1707$, $P = 0.3173$, supplementary fig. S4, Supplementary Material online).

Discussion

DNA Methylation Is Associated with Gene Copy Number in *A. mellifera*

We analyzed the levels of gene body methylation in singleton and duplicate genes of the honeybee in order to understand whether epigenetic factors may be associated with the fate of paralogs. We found a significant difference between the mean CpG methylation levels of singleton and duplicate genes (fig. 1). Singletons consistently had higher levels of gene body methylation than duplicates, which were themselves more highly methylated than genes in larger gene families. This striking result suggests that gene body methylation may be relevant to gene duplication in insects.

Prior studies in methylation of duplicate genes have mainly been limited to vertebrate and plant species. For example, duplicated genes showed higher levels of DNA methylation of promoters than singletons (Chang and Liao 2012; Xu et al. 2018). These differences in promoter methylation were viewed as potentially indicating that DNA methylation was involved in gene dosage rebalancing after a duplication event. In vertebrate and plant systems, DNA methylation of the promoter region has been causatively linked to decreased gene expression via gene silencing (Weber et al. 2007; Baylin and Jones 2011; Lee and Chen 2001). The findings of differences in promoter methylation of duplicate and singleton genes in vertebrates and plants demonstrate a possible epigenetic role in the regulation of paralogs postduplication in these taxa.

Chang and Liao (2012) also investigated patterns of gene body methylation in a vertebrate system and found higher gene body methylation in duplicates than singletons. This previous study focused on DNA methylation in vertebrate genomes, which show very different methylation patterns than insects. Thus, the contrasting results could arise from the distinct epigenetic systems in vertebrate and insect taxa. Nevertheless, the presence of differential gene body methylation in singleton and duplicate genes of *A. mellifera*, akin to the differential promoter methylation in plant and vertebrates, supports the idea of gene body methylation being associated with duplicate gene evolution.

We investigated the relative divergence in gene body methylation between paralogs in order to determine whether they were more- or less-similarly methylated than other pairs of genes. Our results indicated that duplicate genes tended to show relatively similar patterns of DNA methylation. This similarity has been shown to exist previously for other epigenetic marks such as histone modifications in yeast, where duplicate genes share more common promoter and open-reading-frame histone code patterns than random singleton pairs (Zou et al. 2012). Because paralogous genes are—by nature of gene duplication—likely to be more similar in sequence than random pairs of genes, the encoded heritable gene body methylation profile of duplicate genes is also likely to be similar.

Changes to the methylation profile, however, could have consequences on divergence of paralog function. DNA methylation of promoters in vertebrate systems has been suggested to be an important factor in the initial divergence in paralogs through repressive effects on gene expression (Rodin and Riggs 2003; Fang et al. 2018). Specifically, promoter methylation could mask one paralog from selection by, in essence, turning the gene off. Under this model, the masked paralog could then accumulate mutations advantageous to its functional divergence. However, methylation of gene bodies in insects is associated with increased and constitutive expression rather than the silencing of expression associated with promoter methylation. Thus, the previously developed models may not directly apply to insect systems, and a new interpretation is necessary to understand the methylation of duplicates in other biological systems.

Caste-Biased Expression of Genes Depends on Gene Copy Number

Previous studies have shown that singleton genes in *A. mellifera* have lower levels of caste- and tissue-biased expression than duplicate genes (Chau and Goodisman 2017). Moreover, singletons consistently have higher levels of gene body methylation, indicating that genes with higher DNA methylation levels are more likely to serve as the highly conserved housekeeping genes (Elango et al. 2009; Hunt et al. 2010; Bonasio et al. 2012; Glastad et al. 2013). Our data reinforce these results by demonstrating strong associations between gene duplication status and biased gene expression. In particular, duplicates tended to display biased gene expression between castes.

Additionally, we found a strong association between caste-biased gene expression and DNA methylation for singleton and duplicate genes in *A. mellifera*. This general correlation between gene body methylation and gene expression bias has been uncovered many times previously in insects (Kucharski et al. 2008; Foret et al. 2009; Lyko et al. 2010; Herb et al. 2012; Li-Byarlay et al. 2013; Glastad et al. 2016; but see Bewick, Sanchez, et al. 2019; Bewick, Zhang, et al. 2019). However, our work investigated the interactions between these previously identified associations within gene body methylation, caste-biased expression, and duplication status. We found that the established associations between caste-biased expression and gene body methylation levels were much stronger in single-copy genes than in duplicated genes (table 1). In other words, duplicate genes and singletons showed different relationships between gene body methylation and expression bias between castes. This suggests that DNA methylation, although differentially applied to single- and multiple-copy genes, does not play a clear causative role in the caste-biased expression of a gene that has been duplicated.

To further probe the importance of gene body methylation to caste-biased patterns of expression, we removed methylation as an explanatory variable. We determined that the association between duplication status and caste-biased expression was much weaker in genes showing low methylation than in genes with high-methylation levels. Duplicated genes tended to show much lower levels of genic methylation on average (fig. 1) and, accordingly, also lacked associations between genic methylation and caste-biased expression (table 1). These analyses demonstrate an association between gene body methylation and biased expression of duplicate genes.

Divergence of Paralogs Reveals Evolution of DNA Methylation and Gene Expression

We compared DNA methylation and gene expression in *A. mellifera* paralogs with single-copy outgroup orthologs from a related bee species, *C. calcarata*. We found that levels of gene expression were correlated to levels of DNA methylation in *A. mellifera*, a result which has been noted previously (e.g., Kucharski et al. 2008; Foret et al. 2009; Lyko et al. 2010). We also found that levels of DNA methylation for the *C. calcarata* ortholog and *A. mellifera* paralogs were highly correlated. This demonstrates the general stability of gene body DNA methylation patterns across insects (Hunt et al. 2010; Zemach et al. 2010; Sarda et al. 2012; Glastad et al. 2013). Patterns of gene expression were also highly correlated between the outgroup ortholog and each paralogous gene copy, again indicating general conservation of gene expression patterns over time. Interestingly, however, the gene expression levels were not significantly correlated between the two *A. mellifera* paralogous gene copies. This result suggests that the paralogous gene copies have diverged in levels of expression, presumably as some duplicated genes gain novel expression profiles and functions (Holland et al. 1994; Li et al. 2018).

Next, we examined the relative divergence in methylation and expression between paralogs and outgroup orthologs to understand if methylation and expression evolution were associated. We found that relative divergence of gene body methylation was correlated with relative divergence of gene expression for the *A. mellifera* paralogs. These results suggest a connection between gene body methylation evolution and gene expression evolution.

To further investigate this connection between DNA methylation and gene expression, we next compared relative divergence of methylation between the *A. mellifera* paralogs and the outgroup ortholog against relative divergence of gene expression between *A. mellifera* paralogs and the ortholog.

We predicted a positive association between these divergence estimates, which would indicate that as methylation divergence increased between orthologous copies so did expression divergence. Surprisingly, however, we uncovered a marginally negative association between these methylation and divergence estimates, suggesting that gene body methylation and gene expression were not necessarily always directly linked (Roudier et al. 2009; Bewick et al. 2016; Bewick, Sanchez, et al. 2019; Bewick, Zhang, et al. 2019).

Finally, we investigated the distribution of relative divergence estimates between *C. calcarata* orthologs and *A. mellifera* paralogs. We reasoned that the distribution of these estimates should be more or less Gaussian if evolution of gene expression and DNA methylation occurred predominantly through random processes. However, we found that the distributions of relative divergence estimates for both gene expression and DNA methylation differed significantly from a normal distribution. Interestingly, these differences came in the form of an excess of values around the mean, rather than an excess of values in the tails (fig. 2). We suggest that the lack of genes at intermediate values could indicate selective effects removing duplicated genes that diverge at marginal levels of methylation or expression. Instead, selection may act to preserve genes showing conserved or extreme patterns of genic methylation or gene expression.

Patterns of DNA Methylation between Duplicate Genes Change over Time

We investigated whether gene body methylation divergence between paralogs changed over time. We found that putatively younger pairs of duplicate genes generally had higher levels of DNA methylation divergence than older pairs (fig. 3). This suggests that the DNA methylation level of a duplicated gene becomes more similar to its sister paralog as the pair ages. This result is surprising, since one might expect patterns of methylation to diverge as a pair of duplicate genes ages.

A previous study of human duplicate genes found no significant correlation between gene body methylation divergence and evolutionary time but did uncover a positive association between promoter methylation divergence and evolutionary time (Keller and Yi 2014). Moreover, studies of DNA methylation in *Arabidopsis* demonstrated a positive correlation between gene body methylation divergence and sequence divergence of paralogs (Wang et al. 2014), possibly indicating the evolution of differential expression between paralogs in *Arabidopsis*. The observed differences in patterns of paralog methylation divergence in different species could reflect the distinct methylation contexts in which they are seen; insect, vertebrate, and plant systems all show differences in levels and targets of DNA methylation, which may affect patterns of DNA methylation divergence.

We also investigated how overall DNA methylation level of paralogs changed over time. We found that older pairs of duplicate genes displayed lower overall levels of gene body methylation. That is, it appears that duplicates lost methylation as they aged. Zhong et al. (2016) demonstrated that promotor methylation decreased with evolutionary time (dS) in zebrafish; in contrast, gene body methylation was significantly higher in older duplicates than in younger duplicates. These differences in methylation patterns between the honeybee and zebrafish could be a result of the different methylation systems in the two taxa. The decrease in overall methylation of duplicates over time could also be a factor in the observed differences in methylation between singleton and duplicate genes. Singleton genes, as defined and measured in this study, are likely to be older than duplicated genes. The interplay between divergence time and DNA methylation adds a layer of complexity to the system, making it likely that the observed differences result from multiple factors.

Methylation Divergence and the Genomic Location of Paralogs

We investigated the methylation levels of genes on the 16 *A. mellifera* chromosomes. We found that most chromosomes showed similar levels of DNA methylation, with only 1 out of 16 chromosomes showing a significant difference in genic methylation level from the majority of the others (supplementary fig. S2, Supplementary Material online). This suggests that any epigenetic “chromosomal effect” on DNA methylation of genes, which could differentially affect the function of genes residing on different chromosomes, is limited and does not account for potential divergences in gene body methylation between nonsyntenic paralogs (fig. 4).

In contrast, we found that duplicated genes did show significantly higher methylation divergence when located on different chromosomes rather than when located on the same chromosome. This result reinforces previous work in *Arabidopsis* and *Oryza* that showed that retrotransposed and dispersed paralogs, which are often found on different chromosomes, had a higher divergence in gene body methylation than tandem duplications, which are typically located on the same chromosome (Wang et al. 2014, 2017). In vertebrates, it has been demonstrated that putatively younger duplicate gene pairs (by measure of synonymous substitution rate ratio) are more likely to be syntenic than older duplicate pairs (Rodin et al. 2005). If younger duplicates in insects also tend to be syntenic, then higher divergence in gene body methylation of nonsyntenic paralogs could be indicative of their evolutionary divergence as the pair ages.

Gene Length and Gene Body Methylation Are Associated with Gene Copy Number

We found that genes with a lower genic methylation level were significantly longer than those with higher methylation. Previous studies also demonstrated a negative correlation between gene length and gene body methylation in *A. mellifera* (Zeng and Yi 2010). The novel result in our study, however, is the degree to which this correlation seemingly differed between singletons and duplicated genes (supplementary fig. S4, Supplementary Material online). Duplicated genes did not show a significant correlation between length and level of DNA methylation when long genes were included in the analysis. The correlation was quite strong, however, in singletons.

Previous studies have shown that genes with lower levels of promoter methylation, and subsequently increased transcription, tend to be shorter than genes with higher levels of promoter methylation (Takuno and Gaut 2012; Zhong et al. 2016). It has been proposed that constitutively expressed and biologically essential housekeeping genes could evolve to be shorter than tissue-biased genes in order to maximize the efficiency of transcription and increase expression (Eisenberg and Levanon 2003; Urrutia and Hurst 2003). *Apis mellifera* housekeeping genes are associated with high levels of gene body methylation and moderate to high levels of expression (Elango et al. 2009; Hunt et al. 2010; Bonasio et al. 2012; Glastad et al. 2013). We found that shorter genes are associated with higher levels of gene body methylation and could therefore be representative of these conserved housekeeping genes. Duplicated genes, conversely, would be associated with divergent function and biased expression, as has been shown in *A. mellifera* previously (Chau and Goodisman 2017).

Conclusions

This study revealed that patterns of gene body methylation differed between single-copy genes and duplicated genes. We also discovered associations in duplicate gene methylation in other contexts including gene pair divergence time, location, and length. Our work demonstrates the potential of gene body methylation to affect the regulation and evolution of paralogs. Future studies could investigate whether patterns of DNA methylation differ between duplicates and singletons in other taxa that possess different systems of DNA methylation. Similarly, controlled experimental systems allowing more targeted probing into associations between duplicate gene methylation and gene function will be important. Overall, further research on the effects of epigenetic marks in diverse biological systems will be crucial to understanding the evolution of duplicate genes and the mechanisms of gene regulation.

Materials and Methods

Full details of materials and methods are provided in Supplementary Material online. Gene family information from OrthoDB v9.1 (Zdobnov et al. 2017) was used to identify genes that were found in single and multiple copy in *A. mellifera*. Following identification of gene families, genes with the lowest pairwise dS value in the family were identified as duplicate pairs within those larger family contexts. Whole-genome bisulfite sequencing data from three studies of DNA methylation in the honeybee (Lyko et al. 2010; Herb et al. 2012; Li-Byarlay et al. 2013) were downloaded from the European Nucleotide Archive and processed (Andrews 2010; Leinonen et al. 2011; Martin 2011; Munoz-Torres et al. 2011). Trimmed FastQ files were aligned to the indexed reference using Bismark (Krueger and Andrews 2011), and Bowtie2 (Langmead and Salzberg 2012; supplementary table S6, Supplementary Material online). Methylation calls from Bismark were imported into SeqMonk (Andrews 2020), and reading frames (probes) were created at the level of individual genes (Elsik et al. 2014). Sequenced RNA data were obtained from previous studies that examined gene expression differences in *A. mellifera* between drone, worker, and queen larvae (Ashby et al. 2016) and drone, worker, and queen pupae (Vleurinck et al. 2016).

ANOVA was used to determine if the mean methylation level of genes belonging to families of different sizes differed significantly. χ^2 tests of independence were used to test the association between "duplication status" and "methylation status" of duplicated genes. We then tested if the relative methylation divergence between duplicate gene pairs (Keller and Yi 2014) differed from null expectations using a pooled randomization test. For sequence divergence analyses, multiple protein alignments between duplicate genes were generated using MUSCLE aligner (Edgar 2004). Codon alignments were then created using PAL2NAL (Suyama et al. 2006) and used to calculate synonymous substitution rate ratios between paralogs using the PAML yn00 package (Yang 2007).

OrthoDB v9.1 (Zdobnov et al. 2017) was used to identify *A. mellifera* duplicate genes that had single-copy orthologs in the bee *C. calcarata*. *Ceratina calcarata* de novo genome, methylome, and transcriptome files were obtained from previous studies (Rehan et al. 2014, 2016). Spearman's rank correlations were used to determine whether relationships existed between the methylation/expression levels of the three orthologous genes and to find correlations between relative divergence of percent genic methylation and levels of expression. The distributions in the relative divergence of methylation and expression between *C. calcarata* and each *A. mellifera* paralog were tested against a normal distribution.

Metadata regarding chromosomal location were extracted from OrthoDB gene annotations for downstream analysis using custom perl scripts. ANOVA was used to determine whether the 16 *A. mellifera* chromosomes differed in their mean cytosine methylation percentage. ANOVA was used to determine if the divergence of syntenic duplicates differed significantly from the divergence of nonsyntenic duplicates. A randomization trial of pooled singletons was used as a control for this analysis. ANOVA was used to test the difference in the mean gene body methylation level of each length bin.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgment

This work was supported by the Elizabeth Smithgall Watts Fund.

References

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed April 12, 2020.

Andrews S. 2020. Seqmonk: a tool to visualise and analyse high throughput mapped sequence data. Available from: https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/. Accessed April 12, 2020.

Ashby R, Foret S, Searle I, Maleszka R. 2016. MicroRNAs in honey bee caste determination. *Sci Rep*. 6(1):18794.

Baylin SB, Jones PA. 2011. A decade of exploring the cancer epigenome—biological and translational implications. *Nat Rev Cancer* 11(10):726–734.

Berger SL, Kouzarides T, Shiekhattar R, Shilatifard A. 2009. An operational definition of epigenetics. *Genes Dev*. 23(7):781–783.

Bewick AJ, Ji L, Niederhuth CE, Willing EM, Hofmeister BT, Shi X, Wang L, Lu Z, Rohr NA, Hartwig B, et al. 2016. On the origin and evolutionary consequences of gene body DNA methylation. *Proc Natl Acad Sci U S A*. 113(32):9111–9116.

Bewick AJ, Sanchez Z, Mckinney EC, Moore AJ, Moore PJ, Schmitz RJ. 2019. Dnmt1 is essential for egg production and embryo viability in the large milkweed bug, *Oncopeltus fasciatus*. *Epigenet Chromatin* 12(1):6.

Bewick AJ, Zhang Y, Wendte JM, Zhang X, Schmitz RJ. 2019. Evolutionary and experimental loss of gene body methylation and its consequence to gene expression. *G3 (Bethesda)*. 9:2441–2445.

Bird A. 2002. DNA methylation patterns and epigenetic memory. *Genes Dev*. 16(1):6–21.

Bonasio R, Li Q, Lian J, Mutti NS, Jin L, Zhao H, Zhang P, Wen P, Xiang H, Ding Y, et al. 2012. Genome-wide and caste-specific DNA methylomes of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Curr Biol*. 22(19):1755–1764.

Bonduriansky R, Chenoweth SF. 2009. Intralocus sexual conflict. *Trends Ecol Evol*. 24(5):280–288.

Chang AY, Liao BY. 2012. DNA methylation rebalances gene dosage after mammalian gene duplications. *Mol Biol Evol*. 29(1):133–144.

Chau LM, Goodisman M. 2017. Gene duplication and the evolution of phenotypic diversity in insect societies. *Evolution* 71(12):2871–2884.

Conant GC, Wolfe KH. 2008. Probabilistic cross-species inference of orthologous genomic regions created by whole-genome duplication in yeast. *Genetics* 179(3):1681–1692.

Connallon T, Clark AG. 2011. The resolution of sexual antagonism by gene duplication. *Genetics* 187(3):919–937.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32(5):1792–1797.

Eisenberg E, Levanon EY. 2003. Human housekeeping genes are compact. *Trends Genet*. 19(7):362–365.

Elango N, Hunt BG, Goodisman MAD, Yi SV. 2009. DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, *Apis mellifera*. *Proc Natl Acad Sci U S A*. 106(27):11206–11211.

Ellegren H, Parsch J. 2007. The evolution of sex-biased genes and sex-biased gene expression. *Nat Rev Genet*. 8(9):689–698.

Elsik CG, Worley KC, Bennett AK, Beye M, Camara F, Childers CP, de Graaf DC, Debyser G, Deng J, Devreese B, et al. 2014. Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC Genomics*. 15(1):86.

Fang C, Zou C, Fu Y, Li J, Li Y, Ma Y, Zhao S, Li C. 2018. DNA methylation changes and evolution of RNA-based duplication in *Sus scrofa*: based on a two-step strategy. *Epigenomics* 10(2):199–218.

Feng S, Cokus SJ, Zhang X, Chen P-Y, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, et al. 2010. Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A*. 107(19):8689–8694.

Force A, Lynch M, Pickett FB, Amores A, Yan Y-L, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151(4):1531–1545.

Foret S, Kucharski R, Pittelkow Y, Lockett GA, Maleszka R. 2009. Epigenetic regulation of the honey bee transcriptome: unravelling the nature of methylated genes. *BMC Genomics*. 10(1):472.

Gallach M, Betran E. 2011. Intralocus sexual conflict resolved through gene duplication. *Trends Ecol Evol*. 26(5):222–228.

Glastad K, Hunt B, Goodisman M. 2013. Evidence of a conserved functional role for DNA methylation in termites. *Insect Mol Biol*. 22(2):143–154.

Glastad KM, Gokhale K, Liebig J, Goodisman MA. 2016. The caste- and sex-specific DNA methylome of the termite *Zootermopsis nevadensis*. *Sci Rep*. 6(1):37110.

Herb BR, Wolschin F, Hansen KD, Aryee MJ, Langmead B, Irizarry R, Amdam GV, Feinberg AP. 2012. Reversible switching between epigenetic states in honeybee behavioral subcastes. *Nat Neurosci*. 15(10):1371–1373.

Holland PW, Garcia-Fernàndez J, Williams N, Sidow A. 1994. Gene duplications and the origins of vertebrate development. *Dev Suppl*. 1994:125–133.

Hunt BG, Brisson JA, Yi SV, Goodisman MA. 2010. Functional conservation of DNA methylation in the pea aphid and the honeybee. *Genome Biol Evol*. 2:719–728.

Jones PA. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet*. 13(7):484–492.

Keller TE, Yi SV. 2014. DNA methylation and evolution of duplicate genes. *Proc Natl Acad Sci U S A*. 111(16):5932–5937.

Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27(11):1571–1572.

Kucharski R, Maleszka J, Foret S, Maleszka R. 2008. Nutritional control of reproductive status in honeybees via DNA methylation. *Science* 319(5871):1827–1830.

Kucharski R, Maleszka J, Maleszka R. 2016. A possible role of DNA methylation in functional divergence of a fast evolving duplicate gene encoding odorant binding protein 11 in the honeybee. *Proc R Soc B* 283(1833):20160558.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 9(4):357–359.

Lee H-S, Chen ZJ. 2001. Protein-coding genes are epigenetically regulated in *Arabidopsis* polyploids. *Proc Natl Acad Sci U S A*. 98(12):6753–6758.

Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tarraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, et al. 2011. The European nucleotide archive. *Nucleic Acids Res*. 39(Database):D28–D31.

Li Z, Tiley GP, Galuska SR, Reardon CR, Kidder TI, Rundell RJ, Barker MS. 2018. Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proc Natl Acad Sci U S A*. 115(18):4713–4718.

Li-Byarlay H, Li Y, Stroud H, Feng S, Newman TC, Kaneda M, Hou KK, Worley KC, Elsik CG, Wickline SA, et al. 2013. RNA interference knockdown of DNA methyl-transferase 3 affects gene alternative splicing in the honey bee. *Proc Natl Acad Sci U S A*. 110(31):12750–12755.

Lorincz MC, Dickerson DR, Schmitt M, Groudine M. 2004. Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nat Struct Mol Biol*. 11(11):1068–1075.

Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. 2010. Regulation of alternative splicing by histone modifications. *Science* 327(5968):996–1000.

Lyko F, Foret S, Kucharski R, Wolf S, Falckenhayn C, Maleszka R. 2010. The honey bee epigenomes: differential methylation of brain DNA in queens and workers. *PLoS Biol*. 8(11):e1000506.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290(5494):1151–1155.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J*. 17(1):10–12.

Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, et al. 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* 466(7303):253–257.

Munoz-Torres MC, Reese JT, Childers CP, Bennett AK, Sundaram JP, Childs KL, Anzola JM, Milshina N, Elsik CG. 2011. Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera. *Nucleic Acids Res*. 39(Database):D658–D662.

Nei M. 1969. Gene duplication and nucleotide substitution in evolution. *Nature* 221(5175):40–42.

Niederhuth CE, Bewick AJ, Ji L, Alabady MS, Kim KD, Li Q, Rohr NA, Rambani A, Burke JM, Udall JA, et al. 2016. Widespread natural variation of DNA methylation within angiosperms. *Genome Biol*. 17(1):194.

Normark BB. 2003. The evolution of alternative genetic systems in insects. *Annu Rev Entomol*. 48(1):397–423.

Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.

Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annu Rev Genet*. 34(1):401–437.

Pennell TM, Holman L, Morrow EH, Field J. 2018. Building a new research framework for social evolution: intralocus caste antagonism. *Biol Rev*. 93(2):1251–1268.

Qian W, Liao BY, Chang AY, Zhang J. 2010. Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends Genet*. 26(10):425–430.

Ramirez-Gonzalez RH, Borrill P, Lang D, Harrington SA, Brinton J, Venturini L, Davey M, Jacobs J, van Ex F, Pasha A, et al. 2018. The transcriptional landscape of polyploid wheat. *Science* 361(6403):eaar6089.

Rehan SM, Berens AJ, Toth AL. 2014. At the brink of eusociality: transcriptomic correlates of worker behaviour in a small carpenter bee. *BMC Evol Biol*. 14(1):260.

Rehan SM, Glastad KM, Lawson SP, Hunt BG. 2016. The genome and methylome of a subsocial small carpenter bee, *Ceratina calcarata*. *Genome Biol Evol*. 8(5):1401–1410.

Rodin SN, Parkhomchuk DV, Rodin AS, Holmquist GP, Riggs AD. 2005. Repositioning-dependent fate of duplicate genes. *DNA Cell Biol*. 24(9):529–542.

Rodin SN, Riggs AD. 2003. Epigenetic silencing may aid evolution by gene duplication. *J Mol Evol*. 56(6):718–729.

Roudier F, Teixeira FK, Colot V. 2009. Chromatin indexing in *Arabidopsis*: an epigenomic tale of tails and more. *Trends Genet*. 25(11):511–517.

Sarda S, Zeng J, Hunt BG, Yi SV. 2012. The evolution of invertebrate gene body methylation. *Mol Biol Evol*. 29(8):1907–1916.

Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, Oberdoerffer P, Sandberg R, Oberdoerffer S. 2011. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* 479(7371):74–79.

Simmen MW, Leitgeb S, Charlton J, Jones SJM, Harris BR, Clark VH, Bird A. 1999. Nonmethylated transposable elements and methylated genes in a chordate genome. *Science* 283(5405):1164–1167.

Simpson SJ, Sword GA, Lo N. 2011. Polyphenism in insects. *Curr Biol*. 21(18):R738–R749.

Stephens S. 1951. Possible significance of duplication in evolution. *Adv Genet*. 4:247–265.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*. 34(Web Server):W609–W612.

Suzuki MM, Bird A. 2008. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet*. 9(6):465–476.

Takuno S, Gaut BS. 2012. Body-methylated genes in *Arabidopsis thaliana* are functionally important and evolve slowly. *Mol Biol Evol*. 29(1):219–227.

Urrutia AO, Hurst LD. 2003. The signature of selection mediated by expression on human genes. *Genome Res*. 13(10):2260–2264.

Vleurinck C, Raub S, Sturgill D, Oliver B, Beye M. 2016. Linking genes and brain development of honeybee workers: a whole-transcriptome approach. *PLoS One* 11(8):e0157980.

Wang J, Marowsky NC, Fan C. 2014. Divergence of gene body DNA methylation and evolution of plant duplicate genes. *PLoS One* 9(10):e110357.

Wang X, Zhang Z, Fu T, Hu L, Xu C, Gong L, Wendel JF, Liu B. 2017. Gene-body CG methylation and divergent expression of duplicate genes in rice. *Sci Rep*. 7(1):2675.

Wang Y, Jorda M, Jones PL, Maleszka R, Ling X, Robertson HM, Mizzen CA, Peinado MA, Robinson GE. 2006. Functional CpG methylation system in a social insect. *Science* 314(5799):645–647.

Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, Rebhan M, Schubeler D. 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet*. 39(4):457–466.

Wheeler DE. 1986. Developmental and physiological determinants of caste in social Hymenoptera: evolutionary implications. *Am Nat*. 128(1):13–34.

Xu C, Nadon BD, Kim KD, Jackson SA. 2018. Genetic and epigenetic divergence of duplicate genes in two legume species. *Plant Cell Environ*. 41(9):2033–2044.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24(8):1586–1591.

Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simao FA, Ioannidis P, Seppey M, Loetscher A, Kriventseva EV. 2017. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res*. 45(D1):D744–D749.

Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-side evolutionary analysis of eukaryotic DNA methylation. *Science* 328(5980):916–919.

Zeng J, Yi SV. 2010. DNA methylation and genome evolution in honeybee: gene length, expression, functional enrichment covary with the evolutionary signature of DNA methylation. *Genome Biol Evol*. 2:770–780.

Zhong Z, Du K, Yu Q, Zhang YE, He S. 2016. Divergent DNA methylation provides insights into the evolution of duplicate genes in zebrafish. *G3 (Bethesda)* 6:3581–3591.

Zilberman D. 2017. An evolutionary case for functional gene body methylation in plants and animals. *Genome Biol*. 18(1):87.

Zilberman D, Henikoff S. 2007. Genome-wide analysis of DNA methylation patterns. *Development* 134(22):3959–3965.

Zou Y, Su Z, Huang W, Gu X. 2012. Histone modification pattern evolution after yeast gene duplication. *BMC Evol Biol*. 12(1):111.

Supporting Information for:

Gene duplication in the honeybee: Patterns of DNA methylation, gene expression, and genomic environment

Carl J. Dyson and Michael A. D. Goodisman

This Supporting Information file includes:

Supplementary Figures S1-S4

Supplementary Materials and Methods

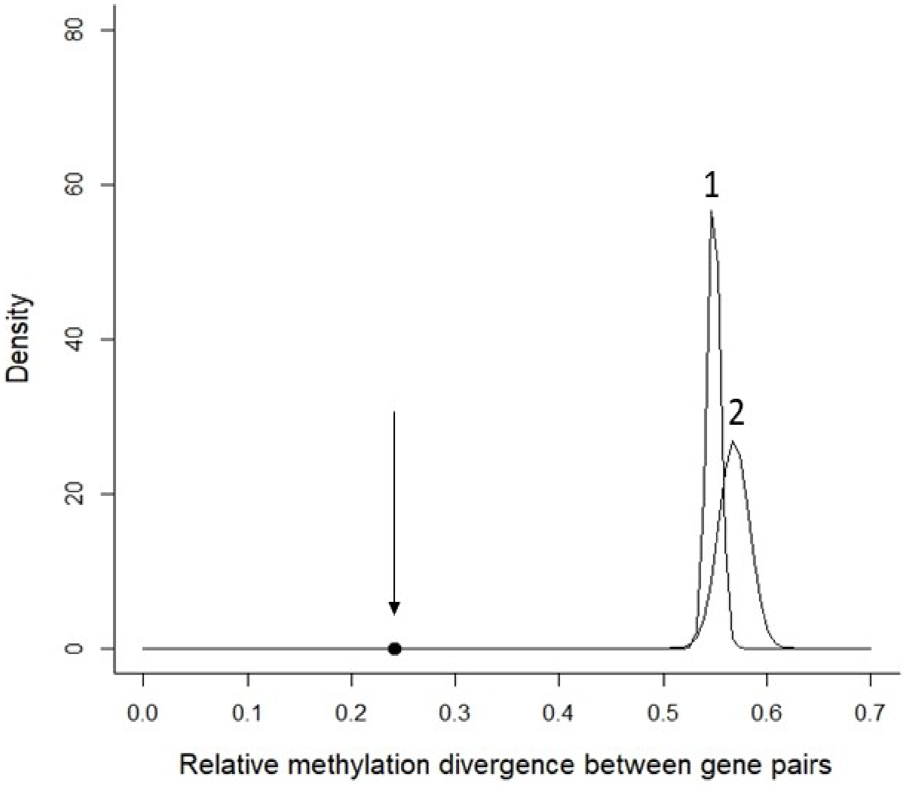

Fig. S1. Null distributions of relative divergence in methylation for 10,000 randomly paired (1) singletons and (2) duplicates. Arrow indicates actual value of relative divergence (D_r) in methylation between duplicate genes in *Apis mellifera*. The true mean value differed significantly from the distributions of randomly generated mean values ($p < 0.0001$).

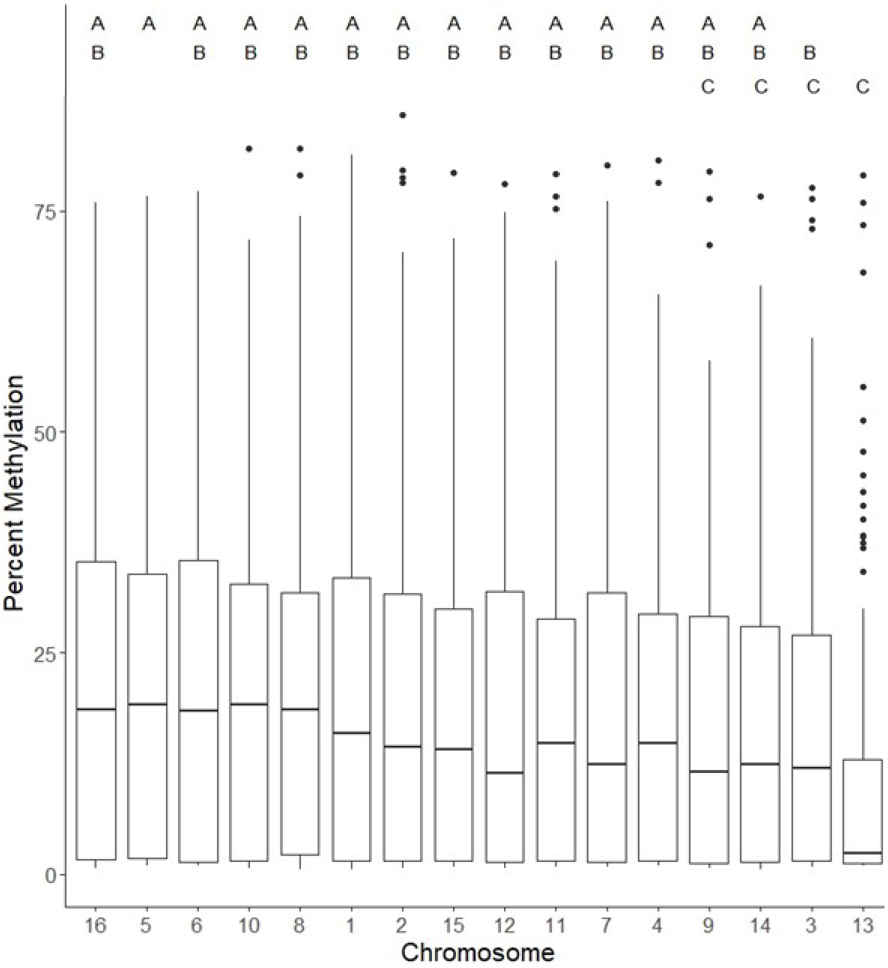

Fig. S2. Percentage of CpG methylation for genes on the 16 chromosomes, ordered from highest to lowest mean percent methylation. The genes on chromosomes denoted with different letters differed significantly in mean percent methylation. Box represents first quartile, median, and third quartile values, while whiskers represent values within 1.5 x the interquartile range, and points represent outliers.

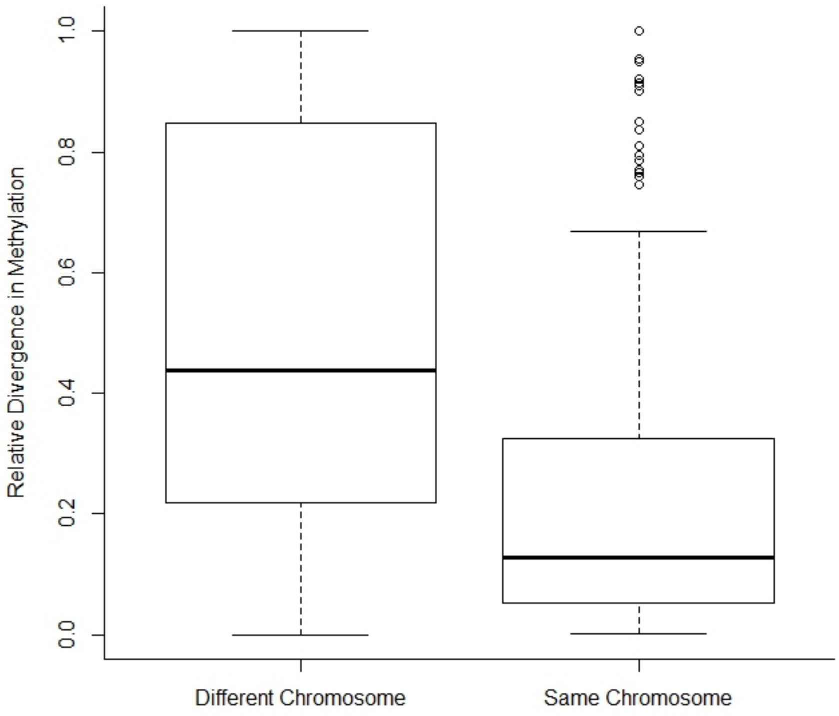

Fig. S3. Duplicated genes on the same chromosome show significantly lower levels of methylation divergence than those on different chromosomes ($p < 0.0001$). Boxes provide first quartile, median, and third quartile values, while whiskers represent values within 1.5 x the interquartile range, and points represent outliers.

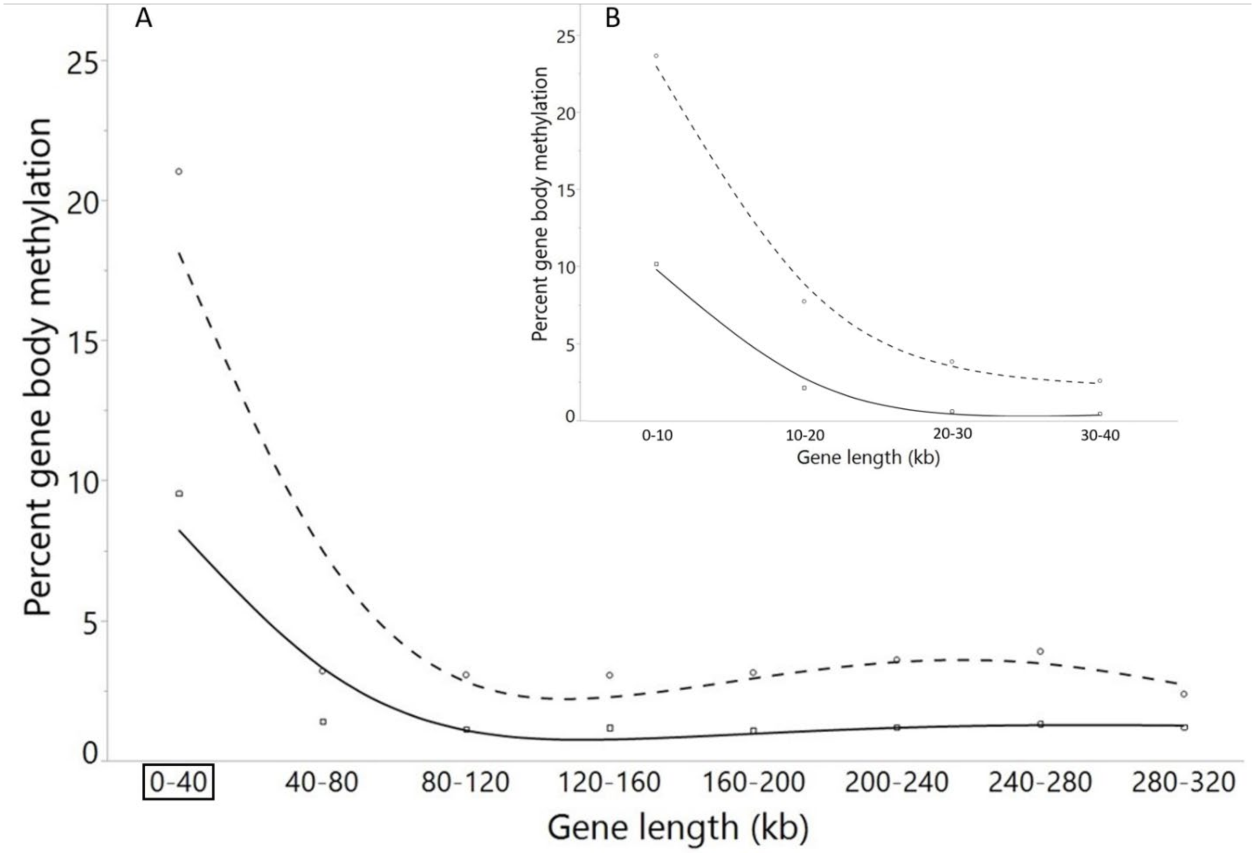

Fig. S4. (A) Relationship between gene body methylation and gene body length in singleton (○, dashed line) and duplicate (□, solid line) genes. Longer singletons and duplicate genes tended to have significantly lower levels of DNA methylation. Points represent mean gene body methylation level for genes in each bin. Lines represent cubic spline fit with lambda (smoothing parameter) of 0.05. ANOVA of means shows that singletons vary significantly in mean methylation across bins ($p < 0.0001$), while duplicates do not ($p = 0.3173$). (B) Relationship between gene length and percent genic methylation of shorter genes (0-40kb) shows a significant negative association in both singletons ($p < 0.0001$) and duplicates ($p = 0.0008$).

Supplemental Materials and Methods

Gene family and DNA methylation datasets

Gene family information from OrthoDB v9.1 (Zdobnov, et al. 2017) was used to identify genes that were found in single- and multiple-copy in *A. mellifera* based off orthologous genes in related taxa (Chau and Goodisman 2017). The set of duplicated genes obtained consisted of all multiple-copy, functional genes in *A. mellifera,* including genes that underwent a duplication event following divergence from other insect species within Apoidea. Following identification of gene families, genes with the lowest pairwise dS value in the family were identified as duplicate pairs within those larger family contexts.

Whole-genome bisulfite sequencing data from three studies of DNA methylation in the honeybee were downloaded from the European Nucleotide Archive (Leinonen, et al. 2010). The three studies focused on (1) DNA methylation differences between queen and worker brains (Lyko F. 2010), (2) the effects of RNA-i knockdown of DNMT3 in workers (Li-Byarlay, et al. 2013), and (3) epigenetic state reversal between worker subclasses (Herb, et al. 2012). In total, 16 datasets from these studies were used for further analyses.

All bisulfite samples were analyzed with FastQC (Andrews 2010) to determine the need for custom trimming parameters. Sequences were trimmed using Trim Galore! (Krueger 2015) and primers and extraneous sequence repeats were removed using Cutadapt (Martin 2011). Trimming was completed with parameters -q 20 -e 0.1 -l 20 (representing a Phred quality score cutoff of 20, a maximum error rate of 0.1, and a minimum post-trim read length of 20bp). The reference genome used for read alignment was *A. mellifera* genome assembly Amel_4.5, downloaded from the Beebase Hymenoptera Genome Database (Munoz-Torres, et al. 2010) and converted to an indexed reference using Bismark Genome Preparation (Krueger and Andrews 2011) with default parameters.

Trimmed FastQ files were aligned to the indexed reference using Bismark and Bowtie2 (Langmead and Salzberg 2012). The alignment was performed using the most stringent parameters for allowed mismatches and seed length. In addition, the --ambiguous tag was applied in order to discard any reads with more than one unique alignment. This alignment setting was included to eliminate multimapping of reads which mapped to multiple areas of the reference genome with the same quality of alignment, which may be an issue when mapping duplicated genes with high sequence similarity. Due to the stringency of alignment parameters, this led to less than 5% of reads needing removal due to non-unique alignments. The analyses were also tested with these reads included, as a control against the data being skewed from this step. CpG methylation calls from the alignment were then extracted using Bismark Methylation Extractor (see supplemental table S6 for data file and alignment information).

Methylation calls from Bismark were imported into SeqMonk (Andrews) for visualization and quantification. Reading frames (probes) were created at the level of individual genes, as defined by the *A. mellifera* Official Gene Set Version 3.2 (Elsik, et al. 2014). Read coverage outliers were removed using SeqMonk following developer recommendations to avoid skewing of the results by data of questionable quality. The Seqmonk Bisulphite Methylation Pipeline was used to calculate the mean percentage of bisulfite-converted CpG sites out of all CpG dinucleotides for each gene.

The correlations between DNA methylation calls between each of the 16 WGBS datasets were calculated using JMP in order to measure the consistency of the data across the different studies, tissues, and individuals (supplemental table S7). A new dataset was created by finding the mean CpG methylation level for each *A. mellifera* gene across the 16 datasets. This new mean set was used as representative data for all other analyses.

Analyses of DNA methylation in duplicated genes and singletons

ANOVA was used to determine if the mean methylation level of genes belonging to families of different sizes differed significantly. CpG methylation for this analysis was defined as the percentage of CpG dinucleotides that were methylated within each gene. Methylation status of genes ("high" vs "low") was determined by estimating the midpoint of the bimodally distributed level of gene methylation in the honeybee. This midpoint was identified as 5% CpG methylation across all datasets, and this value was thus used as the threshold for determining if a gene showed high or low levels of methylation. χ^2 tests of independence were used to test the association between "duplication status" (singletons vs. duplicates) and "methylation status" (high vs. low) of duplicated genes, and to determine whether either of these variables could be predictive of the other. χ^2 analysis was completed using the "Chi-Square Test Calculator" (Social Science Statistics, 2018).

We next investigated a relative measure of divergence between paralogs. Relative divergence (D_r) was calculated as (V1 – V2)/(V1 + V2) where V1 and V2 were the values for the focal gene copies (Keller and Yi 2014). We tested if the relative methylation divergence between duplicate gene pairs differed from null expectations using a randomization test. Specifically, all duplicated genes were pooled together into a single gene set and two of these genes were then randomly selected to produce new pairs of "pseudoparalogs". This procedure was repeated until all genes in the set had been randomly paired. The divergence in methylation level between the pseudoparalogs was calculated, and a mean was found for all the pseudoparalog pairs in the set. This entire trial was repeated 10,000 times to produce a null distribution of mean methylation divergence for pairs of randomly selected duplicate genes. The actual observed mean methylation D_r between paralogs was then compared to the null distribution of mean D_r of randomly paired genes. The observed value of mean methylation D_r was deemed significant if it fell outside of 95% of the randomly generated mean methylation D_r values. This randomization trial was then replicated using singleton genes randomly paired into pseudoparalogs, with a normal distribution again being created from 10,000 trials.

We next examined the relationship between relative methylation divergence of paralogs and their sequence divergence since duplication. We used custom Bioperl scripts to extract nucleotide sequence information from genome sequencing files. Multiple protein alignments between duplicate genes were generated using MUSCLE aligner (Edgar 2004) for each gene. Codon alignments were then created from aligned protein files and multifasta nucleotide sequences using PAL2NAL (Suyama, et al. 2006). Codon alignments were used to calculate synonymous substitution rate ratios between paralogs using the PAML yn00 package (Yang 2007). Duplicate pairs with a dS value greater than 3 were discarded from analyses to avoid genes saturated with substitutions. Genes were placed into one of three bins dependent on dS value from 0-1, 1-2, or 2-3. ANOVA was used to test the difference in the means of each bin and determine whether there was an association between relative methylation divergence and synonymous substitution rate ratio (dS).

Analyses of DNA methylation and gene expression

Sequenced RNA data was obtained from previous studies that examined gene expression differences in *A. mellifera* between drone, worker, and queen larvae (Ashby, et al. 2016) and drone, worker, and queen pupae (Vleurinck, et al. 2016). The reads from these datasets were quantified in SeqMonk (Andrews) using the RNA-seq pipeline to generate a read count per gene value. Genes were identified as differentially expressed genes (DEGs) using a statistical cutoff of FDR >= 0.05 between their differential raw expression. These genes were further subdivided into singletons and duplicates, as well as "high methylation" and "low methylation" as described previously. χ^2 tests of independence were used to determine whether there were associations between gene methylation status and gene expression bias between castes for singletons and for duplicates. We also conducted sample-size corrected analyses of these statistical tests to control for differences in sample sizes for singleton and duplicate genes. These re-analyses showed the same general associations as those in the full datasets.

Divergence of paralogs from outgroup orthologs

OrthoDB v9.1 (Zdobnov, et al. 2017) was used to identify *A. mellifera* duplicate genes that had single copy orthologs in the bee *C. calcarata*. *C. calcarata* was chosen for this analysis due to its relative close relation to *A. mellifera* and the availability of sequences for its DNA methylation and gene expression. Genes lacking methylation or expression data were excluded from the dataset. *C. calcarata de novo* genome, methylome, and transcriptome files were obtained from previous studies on the carpenter bee species (Rehan, et al. 2014; Rehan, et al. 2016).

C. calcarata BS-seq reads were trimmed for quality and adapters with default parameters using cutadapt (Martin 2011), and aligned to the *C. calcarata* reference genome using Bismark and Bowtie2 (Langmead and Salzberg 2012) as previously described. SeqMonk (Andrews) BS-seq quantification pipeline was used to quantify the percentage of methylated CpG dinucleotides per gene location, as annotated in Rehan et al. (2016). Values were parsed to find methylation data of gene locations in *C. calcarata* that corresponded to *A. mellifera* genes, creating a subset of 92 genes. Spearman's rank correlations were used to determine whether relationships existed between the methylation levels of the three orthologous genes. Nonparametric tests were used to mitigate the effects of the differences in the overall levels of genome methylation that are found across insects.

C. calcarata transcriptome reads were trimmed for quality and adapters with default parameters using cutadapt (Martin 2011) and aligned to the *C. calcarata* reference genome using

Bismark and Bowtie2 (Langmead and Salzberg 2012). SeqMonk (Andrews) RNA-seq quantification pipeline was used to count the raw reads per gene location, as annotated in Rehan, et al. 2016. These values were parsed to find expression raw counts for *C. calcarata* singleton gene locations that are orthologous to *A. mellifera* duplicates from the gene subset previously created using methylation data. This resulted in a subset of 90 genes that had expression data available. Spearman's rank correlations were used to determine whether relationships existed between the expression levels of the orthologous and paralogous genes.

Spearman's rank correlations were used to find correlations between relative divergence of percent genic methylation and levels of expression. The distributions in the relative divergence of methylation and expression between *C. calcarata* and each *A. mellifera* paralog was tested against a normal distribution using a Shapiro – Wilcoxon Goodness of Fit test to determine whether divergence values significantly differed from a normal distribution centered around a mean of 0.

Analyses of DNA methylation and location of paralogs

Metadata regarding chromosomal location and gene start and end sites was extracted from OrthoDB gene annotations for downstream analysis using custom perl scripts. ANOVA was used to determine whether the 16 *A. mellifera* chromosomes differed in their mean cytosine methylation percentage. Duplicated genes were assigned as being syntenic (located on the same chromosome) or non-syntenic (on different chromosomes) depending on whether the paralogs were located within the same linkage group. The chromosomal location of each gene was taken from OrthoDB v9.1 orthologous groups and combined into seventeen bins: sixteen *A. mellifera* linkage groups (chromosomes) and one unassociated group for genes that were not annotated with a chromosome location (unassociated genes were not used in further analyses).

Mean methylation divergence levels were calculated for syntenic and non-syntenic duplicate pairs. We used ANOVA to determine if the divergence of syntenic duplicates differed significantly from the divergence of non-syntenic duplicates. To establish a control group as a reference for the duplicate gene analyses, 734 singletons were paired at random to mirror the sample size of duplicates. These singletons were then assigned as being on the syntenic or non-syntenic based on their linkage group designation from OrthoDB v9.1 orthology. Mean methylation divergence was calculated for these pseudoparalogs. ANOVA was then used to determine whether methylation divergence differed significantly between pseudoparalogs on the same or different chromosomes. This randomization trial was repeated ten times and the mean of the ANOVA results was obtained to produce a control value for comparison with the ANOVA results of the true duplicate pairs.

Analyses of DNA methylation and gene length

Gene length was defined as the nucleotide sequence difference between the start and end sites of a gene, with start and end sites of genes extracted from orthoDB metadata. CpG methylation percentage for each gene was defined as the percentage of CpG dinucleotides within each gene that were methylated. Genes were placed into one of eight bins based on length, with a width of 40kb per bin. ANOVA was used to test the difference in the mean gene body methylation level of each length bin. This test was performed separately using duplicate and singleton genes. Additionally, ANOVA was used to determine if singleton genes and duplicated genes differed in length overall.