DNA methylation and evolution of duplicate genes

Thomas E. Keller and Soojin V. Yi[1]

School of Biology, Georgia Institute of Technology, Atlanta, GA 30332

Edited* by Wen-Hsiung Li, University of Chicago, Chicago, IL, and approved March 12, 2014 (received for review November 14, 2013)

The evolutionary mechanisms underlying duplicate gene maintenance and divergence remain highly debated. Epigenetic modifications, such as DNA methylation, may contribute to duplicate gene evolution by facilitating tissue-specific regulation. However, the role of epigenetic divergence on duplicate gene evolution remains little understood. Here we show, using comprehensive data across 10 diverse human tissues, that DNA methylation plays critical roles in several aspects of duplicate gene evolution. We first demonstrate that duplicate genes are initially heavily methylated, before gradually losing DNA methylation as they age. Within each pair, DNA methylation divergence between duplicate partners increases with evolutionary age. Importantly, tissue-specific DNA methylation of duplicates correlates with tissue-specific expression, implicating DNA methylation as a causative factor for functional divergence of duplicate genes. These patterns are apparent in promoters but not in gene bodies, in accord with the complex relationship between gene-body DNA methylation and transcription. Remarkably, many duplicate gene pairs exhibit consistent division of DNA methylation across multiple, divergent tissues: For the majority (73%) of duplicate gene pairs, one partner is always hypermethylated compared with the other. This is indicative of a common underlying determinant of DNA methylation. The division of DNA methylation is also consistent with their chromatin accessibility profiles. Moreover, at least two sequence motifs known to interact with the Sp1 transcription factor mark promoters of more hypomethylated duplicate partners. These results demonstrate critical roles of DNA methylation, as well as complex interaction between genome and epigenome, on duplicate gene evolution.

gene duplication | tissue specificity | epigenomics | genomics

Gene duplication is a main process generating genomic repertoires for subsequent functional innovation (1, 2). Elucidating the mechanisms by which newly duplicated genes are retained and evolve new functions is one of the most fundamental problems in molecular evolution. Even though significant progresses have been made (3–9), detailed molecular mechanisms of duplicate gene divergence are not satisfactorily resolved (10, 11).

Here we examine a relatively new avenue of research on duplicate gene evolution: the epigenetic divergence of duplicate genes and how this divergence might relate to functional differentiation of duplicate genes. We focus on DNA methylation. It has been proposed that epigenetic mechanisms such as DNA methylation may facilitate evolution by gene duplication (12). Rodin and Riggs (12) have demonstrated by computational modeling that rates of functional diversification such as subfunctionalization and neofunctionalization increase when epigenetic silencing of duplicates is possible. This advantage of epigenetic silencing is particularly significant when effective population sizes are small, as in humans (12).

Some earlier observations are consistent with epigenetic silencing of duplicate genes. For example, multiple copies of transgenes inserted into plant genomes often get silenced by epigenetic mechanisms such as DNA methylation and/or chromatin alteration (13). Lee and Chen (14) demonstrated that gene copies of specific parental origin in an allotetraploid plant are subject to epigenetic silencing. More recently, Chang and Liao (15) examined evolutionary signatures of DNA methylation [normalized CpG content: CpG O/E (16)] in human and mouse genomes. They showed that duplicate genes exhibit stronger evolutionary signatures of DNA methylation and hence are likely to be more heavily methylated than other genes. Thus, theoretical and empirical data indicate that epigenetic silencing of duplicate genes may play an important role in the evolution of duplicate genes.

However, details of how, or whether, duplicate gene evolution and DNA methylation divergence are coupled are not known. For example, do patterns of DNA methylation change with evolutionary time? Does DNA methylation selectively silence young duplicates? Critically, how does tissue-specific DNA methylation of duplicate genes contribute to the functional divergence of duplicate genes? How variable are patterns of duplicate DNA methylation across different tissues? Here we provide answers to these pressing questions by examining complementary patterns of duplicate gene DNA methylation with evolutionary divergence, gene expression, chromatin accessibility, and motif enrichment across multiple tissues in the human genome.

Results

Hypermethylation of Young Duplicates and Decrease of DNA Methylation in Evolutionary Timescale. We first examined the relationship between DNA methylation levels and the evolutionary ages of duplicate genes. Average promoter DNA methylation levels of duplicate genes are significantly negatively correlated with the evolutionary time measured by the number of substitutions at synonymous sites, dS (Pearson's correlation coefficient $r = -0.4$, $P < 2.2 \times 10^{-16}$ for brain; data are presented in Datasets S1 and S2). This relationship for brain is shown in Fig. 1*A* and for other tissues shown in Fig. S1. Similarly, whole-genome bisulfite-sequencing data displayed the same negative correlation (Table S1). In contrast to this pattern in promoters, gene-body DNA methylation does not show a strong relationship with evolutionary time (Fig. 1*B* for brain). In the whole data, gene-body DNA methylation and dS are negatively correlated, but this correlation is extremely weak (Pearson's $r = -0.06$, $P = 0.001$).

Significance

Duplicate genes are essential and ongoing sources of genetic material that evolution can act on, yet new duplicates are under constant risk of being inactivated by mutations and subsequently lost. We show that a common heritable epigenetic modifier, DNA methylation, plays an important role in duplicate gene evolution. DNA methylation clearly distinguishes young and old duplicates, and the differences in DNA methylation of duplicate genes are associated with functional differences in expression. Remarkably, for a majority of duplicate gene pairs, a specific duplicate partner is consistently hypo- or hypermethylated across highly divergent tissues. Our results indicate that epigenetic modifications are intimately involved in the regulation and maintenance of duplicate genes.

Author contributions: T.E.K. and S.V.Y. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]To whom correspondence should be addressed. E-mail: soojinyi@gatech.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1321420111/-/DCSupplemental.

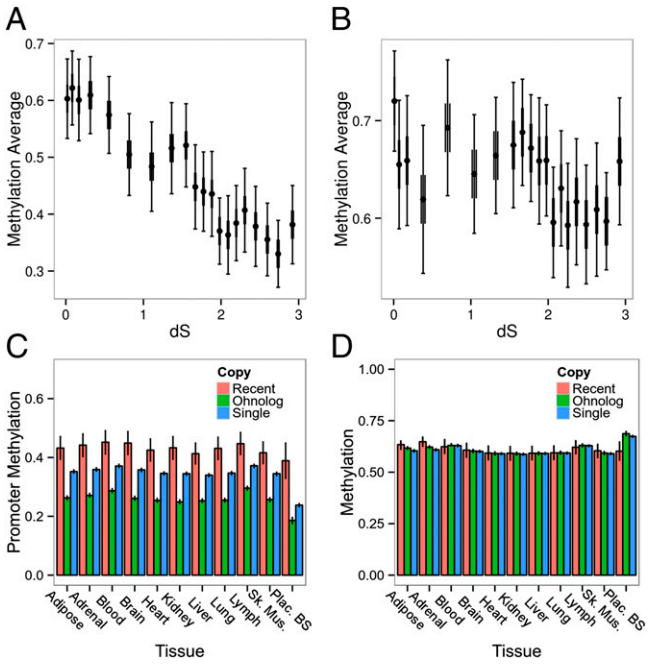

Fig. 1. Contrasting patterns of DNA methylation between young and old duplicates and between promoters and gene bodies. (*A*) The relationship between average promoter DNA methylation levels and evolutionary age are shown for 1,325 gene pairs with dS <3. Promoters of duplicate genes are heavily methylated initially and gradually lose DNA methylation as they age. (*B*) In contrast, gene-body DNA methylation does not exhibit a consistent pattern across evolutionary age. DNA methylation data from brain are divided into 20 evenly sized bins by dS. (*C*) Recent duplicates (since human–macaque divergence, $n = 138$ pairs) are more heavily methylated than single-copy genes ($n = 13{,}871$) as well as ohnologs ($n = 1{,}062$ pairs). (*D*) In contrast, gene-body methylation is similar for all three classes. Error bars represent 95% confidence intervals.

In the whole-genome bisulfite-sequencing data, this correlation is not significant (Table S1). Changing the definition of gene bodies to include only exons yielded similar results (Table S1).

These trends are apparent when we compare methylation levels of "young" duplicate pairs (originated since human–macaque divergence, $n = 138$) to those of "old" duplicate pairs (ohnologs generated by whole-genome duplication at the origin of vertebrates, $n = 1{,}062$). Ohnologs are significantly less methylated than young duplicates in promoters (Fig. 1*C*). Interestingly, ohnologs are even less methylated than singletons (Fig. 1*C*). Again, gene-body DNA methylation does not show significant differences between different types of duplicates and between singletons and duplicates (Fig. 1*D*).

DNA Methylation Divergence of Duplicate Partners Along Evolutionary Time. Above we have demonstrated how DNA methylation levels of duplicate genes change with evolutionary time. We then asked how DNA methylation diverges between duplicate partners. For this purpose we calculated the relative DNA methylation divergence (*Materials and Methods*) between duplicate partners within each pair. We found that the relative DNA methylation divergence and evolutionary age are positively correlated (Pearson's $r = 0.28$, $P < 2.2 \times 10^{-16}$ for brain; Datasets S1 and S2). In other words, young duplicate partners tend to display similar levels of methylation (relative DNA methylation divergence is small) compared with older duplicates (Fig. 2*A* for brain, Fig. S2 for other tissues). We also calculated a tissue-specificity index of DNA methylation, which provides information on the relative strengths of DNA methylation across the 10 tissues (TSMI; *Materials and Methods*). Relative divergence of TSMI also increases with evolutionary time (Pearson's $r = 0.10$, $P = 1 \times 10^{-8}$, Fig. 2*B*).

Differential DNA Methylation of Duplicates Correlates with Gene Expression Divergence. Promoter DNA methylation is causatively linked to gene expression in mammals: It is well-established that hypermethylation of promoters silences downstream gene expression (17). We thus examined whether the heavy DNA methylation of young duplicate promoters is associated with reduced levels of gene expression. Indeed, average expression levels of duplicate genes are significantly positively correlated with the evolutionary age of duplicate genes (Pearson's $r = 0.22$, $P < 2.2 \times 10^{-16}$; Fig. 3*A* for brain). We then asked whether differential DNA methylation of duplicate partners (one copy being hypermethylated compared with the other copy) translates to differential levels of gene expression. For this purpose, we calculated the signed relative divergences of DNA methylation and gene expression divergence. For an arbitrarily assigned gene within a gene pair, its DNA methylation (M) and expression (E) are determined. The signed relative divergence is calculated as (M1 − M2)/(M1 + M2) and (E1 − E2)/(E1 + E2). A negative correlation between these measures indicates that the hypermethylated copy exhibits a lower level of gene expression compared with the hypomethylated copy. Fig. 3*B* depicts the relationship between these two variables in the brain; indeed, relative hypermethylation of a duplicate partner indicates a relative reduction in gene expression (Pearson's $r = -0.35$, $P < 2.2 \times 10^{-16}$). This pattern is consistent across all other tissues (Fig. S3).

Interestingly, the slope of the regression between the relative divergence measures was the steepest in younger duplicates compared with older duplicates. For example, in brain the slope for the youngest quartile of duplicates was –0.51, which is higher than that for all duplicates (the slope is −0.4) or compared with the oldest quartile (the slope is −0.28) (Fig. S4). This pattern was consistent across all tissues, suggesting that differential DNA methylation of younger duplicates may be associated with more pronounced expression differences between them compared with older duplicates. However, we remain cautious because these results are observed with high dS values.

One potential caveat of these analyses is that DNA methylation data and gene expression data are obtained from different samples. The results may be affected by interindividual variability. However, it is remarkable that all tissues exhibit similar patterns. Nevertheless, given this potential sample discrepancy we examined the relationship between methylation divergence and expression divergence from two additional sources. We used (*i*) data from an entirely different expression study [the GNF microarray dataset of gene expression (18)] for the 10 tissues

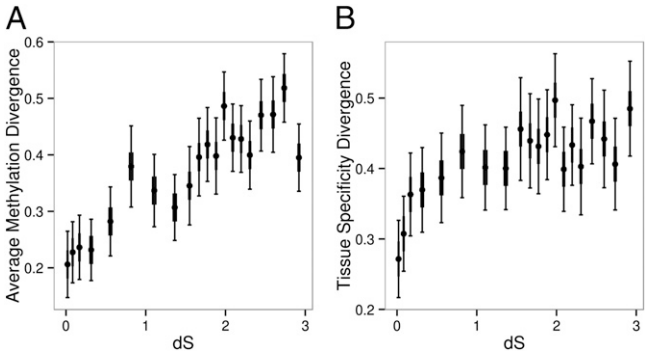

Fig. 2. The relationship between promoter DNA methylation divergence and evolutionary age. (*A*) Recent duplicates have similar levels of methylation, whereas older duplicates are more diverged. Methylation data shown are from brain samples. (*B*) The TSMI calculated across all tissues also increases with evolutionary time. For both panels, data ($n = 1{,}325$ pairs, dS <3) are divided into 20 evenly sized bins by dS.

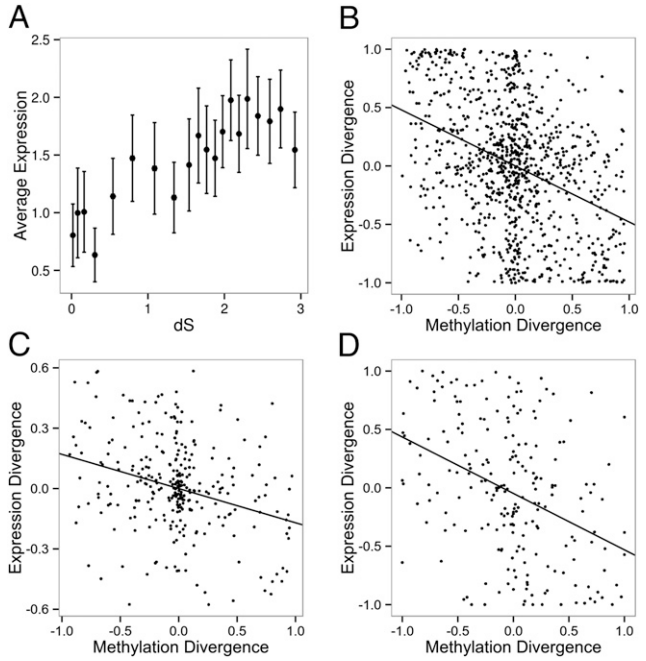

Fig. 3. DNA methylation and expression divergence of duplicate genes. (*A*) Average expression levels of duplicates are lower in young duplicates and increase with evolutionary age (n = 1,298 pairs, divided into 20 evenly sized bins by dS). (*B*) The relationship between signed relative promoter methylation divergence (x axis) and the signed relative expression divergence (y axis). The hypermethylated gene copy is usually less expressed than the hypomethylated copy according to RNA-seq expression data of brain (n = 298 pairs). Likewise, (*C*) brain expression data generated by microarray (n = 1,461 pairs) as well as (*D*) DNA methylation and gene expression generated from the same placenta samples (n = 214 pairs) exhibit the same patterns. The slopes for the regression lines were estimated in R with the "lm" function.

examined and (*ii*) RNA sequencing (RNA-seq) expression data and whole-genome bisulfite-sequencing data from the same placenta sample (19). Analyses of these additional data sets yield very similar results (Pearson's $r = -0.22$, $P = 1.9 \times 10^{-9}$ for GNF data, Fig. 3*C*; Pearson's $r = -0.24$, $P < 2.2 \times 10^{-16}$ for placenta data, Fig. 3*D*). The complementary relationship between DNA methylation divergence and gene expression divergence of duplicate genes seems to be robust against individual variation and experimental platforms.

Majority of Duplicate Genes Exhibit Consistent Patterns of DNA Methylation Divergence Across Tissues. Epigenetic modifications such as DNA methylation are often considered cell type- and individual-specific. According to this idea, we do not necessarily expect gene duplicates to show consistent DNA methylation difference across highly divergent tissues. One partner may be hypermethylated compared with the other in one tissue, but this pattern may be reversed in another tissue. However, DNA methylation is conserved across extremely long evolutionary times (20). Moreover, DNA methylation patterns of specific genomic regions may be determined by sequence characteristics themselves (21).

We thus asked whether duplicate pairs exhibit consistent patterns of DNA methylation divergence across different tissues. Remarkably, we found that for 73% (2,650/3,629) of duplicate gene pairs in the whole-genome dataset, one copy was always consistently less (hypo-) methylated than the other across all 10 tissues. The probability of observing such a number under a model of random (binomial) determination of hypo- and hypermethylation in 10 tissues is extremely small ($P \sim 0$). Although in reality similar cell types may exhibit similar epigenetic profiles, this result nevertheless suggests that there exists a striking commonality in the developmental program that determines the DNA methylation of specific gene duplicates.

Specific DNA Motifs Enriched in Consistently Hypomethylated Promoters. Having identified that the majority of duplicate gene promoters are either consistently hypermethylated or consistently hypomethylated across all 10 tissues, we further examined whether any distinguishing genomic signals exist that separate those two types of promoters. We identified the five most significant motifs in consistently hypermethylated promoters that differ from consistently hypomethylated promoters, and vice versa, using MEME (22). We then used MAST (23) to identify the frequency of these motifs in each promoter set.

None of the motifs identified for hypermethylated promoters was significantly more frequent in hypermethylated promoters after Bonferroni correction (24). In contrast, two motifs identified in hypomethylated promoters were much more likely to occur in the hypomethylated promoters compared with hypermethylated promoters (Fig. 4; Fisher's exact test, $P < 10^{-16}$ for both motifs). Interestingly these motifs contain regions that bind to the zinc finger transcription factor family that includes Sp1, as identified by TOMTOM (Fig. S5) (25). It is particularly notable that these motifs include Sp1 binding sites, because previous experiments have shown that Sp1 binding can prevent local DNA methylation (26). Consequently, the presence of these motifs may serve as one mechanism by which the methylation levels of these promoters have diverged.

Consistently Hypo- and Hypermethylated Duplicate Genes Exhibit Distinctive Chromatin States. We also examined whether members of consistently differentially methylated duplicate pairs exhibit other epigenetic patterns that are common across tissues. We analyzed chromatin accessibility, which measures how "open" specific regions are for regulatory modifications. For this we used DNaseI hypersensitivity data from three brain regions (prefrontal cortex, cerebrum, and cerebellum) (27). Similar data from other examined tissues are currently not available.

We first divided duplicate pairs with DNaseI hypersensitivity data into those that are consistently differentially methylated across all tissues ("consistent pairs," 2,597 pairs) versus those that show fluctuating patterns of differential DNA methylation across tissues ("inconsistent pairs," 894 pairs). We then asked whether the hypermethylated gene copy exhibits more or less open chromatin accessibility compared with the hypomethylated

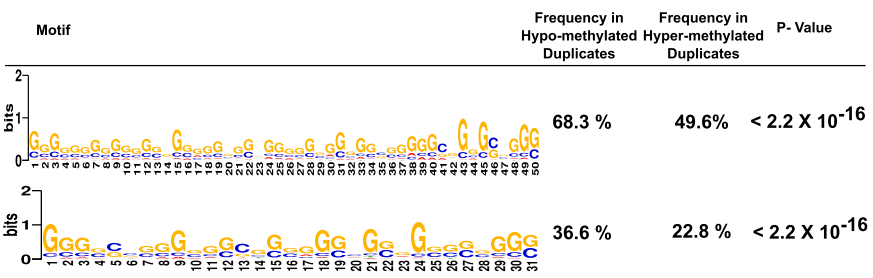

Fig. 4. Consistently hypomethylated duplicates are enriched for specific motifs. Enriched motifs were identified using MEME, and the significance of enrichment was assessed using Fisher's exact test.

copy. For consistent pairs, the hypermethylated copies are much more likely to exhibit less accessible chromatin (Fig. 5*A*, left side, Fisher's exact test, $P < 10^{-16}$). However, there was no difference in the inconsistent pairs (Fig. 5*A*, right side).

We then tested whether there was a quantitative difference in chromatin accessibility between the hyper- and hypomethylated partners in the consistently differentially methylated pairs. The hypomethylated copies displayed significantly higher DNaseI hypersensitivity density compared with the hypermethylated copies (Fig. 5*B*; paired two-sided *t* test, $P < 10^{-16}$). Again, inconsistent pairs had no significant difference between the two copies ($P = 0.64$). The difference in chromatin accessibility was still significant after controlling for the larger sample size of consistent pairs (average *P* value for 1,000 bootstraps of $n = 894$ with replacement $<10^{-16}$). Moreover, we repeated this analysis for the six ENCODE cell lines derived from very different tissues and observed highly similar results (Fig. S6). Together, these results provide strong support for the idea that duplicate gene regulation involves epigenetic modifications sharing common underlying molecular mechanisms across highly different cell types.

Retrogene-Discordant Pairs Show Elevated Methylation and Expression Divergence. The above results begin to reveal that DNA methylation divergence between duplicate partners is partially determined by genomic sequence environments, as well as chromatin environments. In this regard retrogenes, gene duplicates created by the reverse transcription and subsequent insertion of transcribed RNA (28), present an interesting case. Retrogenes may exhibit more pronounced DNA methylation divergence compared with other duplicate genes for the following reasons: First, the transcribed portion of these genes is duplicated onto entirely different genomic sequence environments; second, they are also subject to different chromatin environments. To test this prediction, we identified gene pairs where one copy had multiple exons whereas the second copy consisted of a single exon, according to the NCBI RefSeq database. Those encoded by a single exon are considered as putative retrogenes. We identified 225 gene pairs in our dataset, which we refer to as "retrogene-discordant" gene pairs. Consistent with our prediction, these retrogene-discordant gene pairs are more diverged in promoter methylation compared with nondiscordant gene pairs (Fig. 6; two-sided *t* test, $P = 0.0008$). Interestingly, the putative retrogene in these discordant pairs is more likely to be hypermethylated compared with the gene containing exons (72%, 163/225 pairs, $P = 1.1 \times 10^{-11}$ by binomial test). We then examined whether this pronounced epigenetic divergence leads to more pronounced expression divergence. Indeed, the retrogene-discordant gene pairs display a much stronger negative correlation between methylation and expression divergence (Fig. 6*B*; Pearson's $r = -0.73$, $P = 4.8 \times 10^{-4}$) compared with the whole duplicate gene pairs or those excluding retrogene-discordant gene pairs (Pearson's $r = -0.34$, $P < 2.2 \times 10^{-16}$).

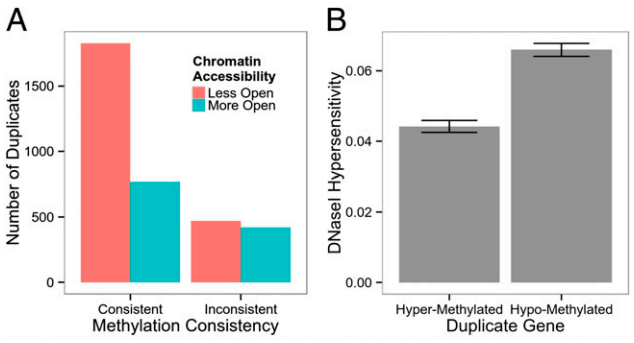

Fig. 5. Consistently hypermethylated duplicate genes have less accessible chromatin. (*A*) The frequencies of hypermethylated partners' being encoded as having less (red) or more (blue) open chromatin. The hypermethylated partner in the consistent pairs (when one gene in a duplicate pair is always hypermethylated than the other across the surveyed tissues) usually has less accessible chromatin compared with the hypomethylated partner. In contrast, the hypermethylated partner in the inconsistent pairs is equally likely to exhibit more or less open chromatin status. (*B*) The hypermethylated partners in consistently differentially methylated pairs have significantly lower DNaseI hypersensitivity compared with the hypomethylated partners.

Discussion

Epigenetic modifications of genomic DNA such as DNA methylation determine how chromatins are compacted and made regionally accessible to regulatory machineries in cellular environments. One aspect of epigenetic modifications that has been of much interest is its pliability in response to external signals. For example, DNA methylation may be altered owing to specific life experience during early development (e.g., ref. 29) and affect active developmental reprogramming. However, evolutionary studies are revealing striking degrees of evolutionary conservation of DNA methylation. For example, gene-body methylation is largely conserved in four highly diverged eukaryotes (20). DNA methylation maps are also strongly conserved between several insect species (30, 31), as well as between rice and *Brachypodium distachyon* despite over 40 million years of divergence (32). Moreover, DNA methylation and histone modifications show unexpected degrees of conservation across distantly related insects (33). These studies suggest that genomic determinants of epigenetic modifications may exist, subject to evolutionary conservation and divergence. To understand the temporal and spatial dynamics of DNA methylation divergence in detail, we analyzed a large number of duplicates spanning different evolutionary ages.

We first show that promoters of young duplicates are hypermethylated in both copies, whereas those of old duplicates are generally hypomethylated. For example, a comparison of duplicates that diverged after human–macaque divergence to those that have diverged at the time of early vertebrate whole-genome duplication demonstrates a clear difference at the level of DNA methylation (Fig. 1). In fact, the latter duplicates are more hypomethylated than singletons. Thus, not all duplicate genes exhibit heavy DNA methylation. Nevertheless, our results are consistent with the "expression reduction model," which states that heavy DNA methylation following the initial duplication event can suppress the expression of duplicate genes, thus providing "buffering" while mutations begin to accumulate (12, 15). Moreover, we show that promoter DNA methylation of duplicates decreases with evolutionary time (Fig. 1 and Fig. S1). Furthermore, relative levels and tissue specificities of DNA methylation between duplicate partners increase with evolutionary time (Fig. 2 and Fig. S2). Together these observations demonstrate that sequence divergence and DNA methylation divergence of duplicate genes are initially coupled.

Moreover, differential DNA methylation covaries with gene expression divergence between duplicates. It was proposed that epigenetic silencing of duplicate genes could facilitate functional divergence of duplicates (12). Indeed, we show that DNA methylation divergence is significantly correlated with gene expression divergence (Fig. 3). Remarkably, this pattern is robust even in unmatched (e.g., expression and methylation measured in different experiments) samples, as well as in matched samples. Although strictly speaking we are only providing evidence of covariation between methylation and expression divergence, the causative relationship between promoter DNA methylation and gene expression has been well established (e.g., ref. 17). Consequently, our study gives strong support to the idea that epigenetic divergence of duplicate genes affects gene expression and, ultimately, functional divergence of duplicate genes.

It is interesting that gene-body DNA methylation does not show a discernible relationship with evolutionary age compared with promoter methylation. This observation is in accord with previous studies that find consistently high methylation in mammalian gene bodies (34, 35). A recent study in rice, however, found that gene-body methylation and divergence is associated

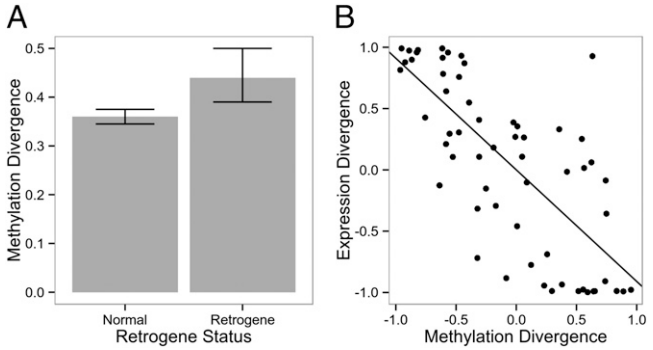

Fig. 6. (*A*) Retrogene-discordant duplicate gene pairs have more highly diverged promoter methylation than other duplicate gene pairs. Error bars indicate 95% confidence intervals. (*B*) The relationship between the signed relative methylation divergence (*x* axis) and the signed relative expression divergence (*y* axis) in retrogene-discordant duplicate gene pairs ($n = 59$). The slope for the regression line was estimated in R using the "lm" function.

with their evolutionary age (36). In addition, duplicates generated by whole-genome duplication versus those generated by small-scale duplications displayed different relationships with respect to DNA methylation and evolutionary divergence in rice (36). In our data we do not observe any qualitative difference in ohnologs beyond the aforementioned differences in DNA methylation levels. These discrepancies may be due to the different patterns and functional consequences of genomic DNA methylation between plants and mammals. In plant genomes, gene-body methylation is the main mode of DNA methylation and is associated with transcription levels (32, 37–39). In contrast, in mammals, transcriptional levels are tightly correlated with promoter DNA methylation, whereas gene-body DNA methylation exhibits a complex relationship with expression levels (e.g., ref. 34). Gene-body DNA methylation in mammals may affect other aspects of gene regulation such as alternative splicing or reduction of transcriptional noise (40, 41). The relationships between gene-body DNA methylation and gene expression and how they vary between different taxa remain to be resolved.

One of the most remarkable findings is that duplicate genes show highly consistent patterns of DNA methylation divergence across multiple tissues. For 73% of gene pairs, one copy is always hypomethylated compared with the other. This refutes the idea that DNA methylation of genes is strictly determined in a tissue-by-tissue manner and indicates the presence of a common developmental programming underlying regulation of duplicate genes across divergent cell types. Consequently, we examined whether we can identify other epigenetic modifications that are similarly consistently regulated across multiple tissues, and whether there exist any specific signals at the genomic sequences themselves. Indeed, we show that gene copies that are consistently hyper- versus hypomethylated across different tissues generally exhibit different chromatin status, the hypomethylated copy typically occupying more open chromatins. This pattern holds both within the tissues used in the primary analysis as well as in very different tissues. Moreover, we discovered at least two sequence motifs that are highly overabundant in hypomethylated compared with hypermethylated duplicate partners. These motifs encode specific transcription factor binding sites, bolstering the general connection between DNA-binding factors and regulation of local DNA methylation (e.g., ref. 42). Furthermore, we demonstrate that gene pairs generated by retrotransposition tend to exhibit more pronounced DNA methylation divergence compared with other gene pairs, strengthening the relationship between genomes and epigenomes.

In summary, our study indicates that epigenetic modifications are important facilitators of duplicate gene evolution owing to their effect on functional divergence as well as potential dependence on genomic determinants. However, duplicate gene evolution provides an excellent system to investigate genetic and environmental factors of epigenetic divergence because we can study divergence of duplicate genes in the same genome, free from other confounding factors.

Materials and Methods

Duplicate Gene Pairs. We constructed a dataset of all human duplicates using Fasta36 (43). Briefly, each protein sequence is compared against every other protein sequence in the human genome. Our criteria for whether two genes were considered a gene pair depended on the length of the alignable region, similarly to ref. 44. We considered two genes to form a pair when the E-score was <10 and the identity of the alignable regions was >30% for regions longer than 150 aa; for shorter alignable regions the identity had to exceed $(0.01n + 4.8L^{-0.32[1+exp(L/1{,}000)]})$, where $n = 6$ and L is the length of the alignable region between the proteins (45). After this initial pairing, we created gene families by performing single-linkage clustering until there were no more groups that shared a member. For each gene family, we first aligned the protein sequences using MUSCLE (46). After calculating pairwise dS using the yn00 module in PAML (47), we selected the gene pair with the lowest dS, realigned them with MUSCLE, and obtained a new dS. We repeated this process until the current gene family was exhausted. This procedure produced the most closely related nonoverlapping gene pairs from each gene family (44). Following this, we identified 3,629 duplicate pairs.

In addition, we assembled two sets of gene duplicates with distinctive evolutionary ages. The first is a dataset of "recent" gene duplicates ($n = 138$ pairs) that are duplicated in the human lineage but single-copy in the rhesus macaque. These are thus at most as old as the human–macaque divergence, ~25 Mya (48). The second set of duplicates is the so-called ohnologs, as previously defined (49), generated by whole-genome duplication near the origin of tetrapods. A total of 1,062 pairs had methylation information for both copies.

DNA Methylation Data. We obtained the methylation data for 482,481 CpGs for 10 different tissues from (50), which used the Infinium HumanMethylation450 Beadchip. Each tissue had at least two biological replicates, which were averaged. Methylation levels were calculated using Illumina Methylation Analyzer (IMA), a statistical package developed in R (51). Methylation probes on the Human Methylation450 array were designed such that they unambiguously map to unique regions of the genome.

We defined the promoter methylation level as the average methylation between 1,500 bases upstream of the transcription start site (TSS) and the end of the first exon; the gene body comprised the remainder of the gene region. We include the first exon in our definition of a promoter because recent studies have shown that the transcriptional effect of methylation typically extends to and includes the first exon (52). In addition, first exons exhibit computational and evolutionary properties that are consistent with promoters (53, 54). We also obtained whole-genome bisulfite-sequencing data of the human placenta from ref. 19 and of human prefrontal cortex brain region from ref. 35. Gene expression levels for these tissues were estimated from paired-end RNA-seq data from the Human BodyMap 2.0 (www.ensembl.info/blog/2011/05/24/human-bodymap-2-0-data-from-illumina/). Raw reads were aligned to the GRCh37 genome using Tophat (55). They were then assembled into transcripts and had their abundances estimated using Cufflinks (56) and log-transformed.

Relative DNA Methylation Divergence. We calculated the signed relative divergence between gene pairs as $(M_1 - M_2)/(M_1 + M_2)$, where M_1 and M_2 are the average methylation levels for the first and second gene, respectively. This measure is similar to that in ref. 57 and quantifies the relative methylation difference between a gene pair, normalized by the overall methylation level of the pair.

Tissue Specificity Index of DNA Methylation. To describe tissue-specific patterns of DNA methylation, we modified the tissue specificity index, which has previously been used for gene expression data (58). We calculated the tissue-specific methylation index (TSMI) as

$$\mathrm{TSMI} = \frac{\sum_{i=1}^{n}[\log 2(m_i)/\log_2(m_{max}) - 1]}{n-1},$$

where m_i is the methylation in tissue i and m_{max} is the maximum methylation level for a focal gene across tissues. As calculated this way, TSMI is a positive value; a small TSMI indicates a relatively even distribution of DNA methylation across tissues, and a large TSMI indicates highly tissue-specific patterns of DNA methylation. Divergence in TSMI between gene pairs was calculated as $(\mathrm{TSMI}_1 - \mathrm{TSMI}_2)/(\mathrm{TSMI}_1 + \mathrm{TSMI}_2)$, similarly to the way the relative DNA methylation divergence was calculated.

Chromatin Accessibility. We used DNaseI hypersensitivity data produced by ENCODE (27), from the "Duke DNaseI HS" download section of the ENCODE project (www.encodeproject.org). The brain tissue IDs were Cerebellumoc, Cerebrumfrontaloc, and Frontalcortexoc. The six cell lines used for the secondary analysis were Gm12878, H1hesc, HelaS3, HepG2, HUVEC, and K562. For each promoter region, we first averaged the hypersensitivity signal across a region and then averaged over cell lines.

Motif Enrichment Analysis. DNA motifs that differentiated the promoter regions of consistently hyper- versus hypomethylated duplicate genes were found using MEME (22). Owing to the large number of sequences being analyzed we restricted the promoter region to the 1,000 bases upstream of the TSS. Repeats were first masked using the RepeatMasker (59). We searched for the top five most-significant motifs that discriminated between the two datasets by generating motif positional priors (60).

ACKNOWLEDGMENTS. We thank Isabel Mendizabal and other members of the S.V.Y. laboratory for their suggestions. Two anonymous reviewers provided excellent comments to improve the manuscript. This research is supported by the Abell Faculty Development Fellowship from the School of Biology of Georgia Institute of Technology and National Science Foundation Grants BCS-1317195 and MCB-0950896 (to S.V.Y.).

1. 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.
2. Ohno S (1970) *Evolution by Gene Duplication* (Springer, New York).
3. Crow KD, Wagner GP; SMBE Tri-National Young Investigators (2006) Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005. What is the role of genome duplication in the evolution of complexity and diversity? *Mol Biol Evol* 23(5): 887–892.
4. Gu ZL, Nicolae D, Lu HHS, Li WH (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet* 18(12):609–613.
5. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290(5494):1151–1155.
6. Shiu SH, Byrnes JK, Pan R, Zhang P, Li WH (2006) Role of positive selection in the retention of duplicate genes in mammalian genomes. *Proc Natl Acad Sci USA* 103(7): 2232–2236.
7. Wagner GP, Amemiya C, Ruddle F (2003) Hox cluster duplications and the opportunity for evolutionary novelties. *Proc Natl Acad Sci USA* 100(25):14603–14606.
8. Conrad DF, et al.; Wellcome Trust Case Control Consortium (2010) Origins and functional impact of copy number variation in the human genome. *Nature* 464(7289): 704–712.
9. Chen FC, Chen CJ, Li WH, Chuang TJ (2010) Gene family size conservation is a good indicator of evolutionary rates. *Mol Biol Evol* 27(8):1750–1758.
10. Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Res* 20(10):1313–1326.
11. Innan H, Kondrashov F (2010) The evolution of gene duplications: Classifying and distinguishing between models. *Nat Rev Genet* 11(2):97–108.
12. Rodin SN, Riggs AD (2003) Epigenetic silencing may aid evolution by gene duplication. *J Mol Evol* 56(6):718–729.
13. Flavell RB (1994) Inactivation of gene expression in plants as a consequence of specific sequence duplication. *Proc Natl Acad Sci USA* 91(9):3490–3496.
14. Lee HS, Chen ZJ (2001) Protein-coding genes are epigenetically regulated in Arabidopsis polyploids. *Proc Natl Acad Sci USA* 98(12):6753–6758.
15. Chang AY, Liao BY (2012) DNA methylation rebalances gene dosage after mammalian gene duplications. *Mol Biol Evol* 29(1):133–144.
16. Yi SV, Goodisman MA (2009) Computational approaches for understanding the evolution of DNA methylation in animals. *Epigenetics* 4(8):551–556.
17. Weber M, et al. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* 39(4):457–466.
18. Su AI, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* 101(16):6062–6067.
19. Schroeder DI, et al. (2013) The human placenta methylome. *Proc Natl Acad Sci USA* 110(15):6037–6042.
20. Sarda S, Zeng J, Hunt BG, Yi SV (2012) The evolution of invertebrate gene body methylation. *Mol Biol Evol* 29(8):1907–1916.
21. Lienert F, et al. (2011) Identification of genetic elements that autonomously determine DNA methylation states. *Nat Genet* 43(11):1091–1097.
22. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 2:28–36.
23. Bailey TL, Gribskov M (1998) Combining evidence using p-values: Application to sequence homology searches. *Bioinformatics* 14(1):48–54.
24. Dunn OJ (1961) Multiple comparisons among means. *J Am Stat Assoc* 56:52–64.
25. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS (2007) Quantifying similarity between motifs. *Genome Biol* 8(2):R24.
26. Brandeis M, et al. (1994) Sp1 elements protect a CpG island from de novo methylation. *Nature* 371(6496):435–438.
27. Thurman RE, et al. (2012) The accessible chromatin landscape of the human genome. *Nature* 489(7414):75–82.
28. Hollis GF, Hieter PA, McBride OW, Swan D, Leder P (1982) Processed genes: A dispersed human immunoglobulin gene bearing evidence of RNA-type processing. *Nature* 296(5855):321–325.
29. Weaver IC, et al. (2004) Epigenetic programming by maternal behavior. *Nat Neurosci* 7(8):847–854.
30. 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.
31. 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.
32. Takuno S, Gaut BS (2013) Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proc Natl Acad Sci USA* 110(5):1797–1802.
33. Hunt BG, Glastad KM, Yi SV, Goodisman MAD (2013) Patterning and regulatory associations of DNA methylation are mirrored by histone modifications in insects. *Genome Biol Evol* 5(3):591–598.
34. Jjingo D, Conley AB, Yi SV, Lunyak VV, Jordan IK (2012) On the presence and role of human gene-body DNA methylation. *Oncotarget* 3(4):462–474.
35. Zeng J, et al. (2012) Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. *Am J Hum Genet* 91(3):455–465.
36. Wang Y, Wang X, Lee T-H, Mansoor S, Paterson AH (2013) Gene body methylation shows distinct patterns associated with different gene origins and duplication modes and has a heterogeneous relationship with gene expression in Oryza sativa (rice). *New Phytol* 198(1):274–283.
37. Feng S, et al. (2010) Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci USA* 107(19):8689–8694.
38. Suzuki MM, Bird A (2008) DNA methylation landscapes: Provocative insights from epigenomics. *Nat Rev Genet* 9(6):465–476.
39. Zemach A, McDaniel IE, Silva P, Zilberman D (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* 328(5980):916–919.
40. Huh I, Zeng J, Park T, Yi SV (2013) DNA methylation and transcriptional noise. *Epigenetics Chromatin* 6(1):9.
41. Maunakea AK, et al. (2010) Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* 466(7303):253–257.
42. Stadler MB, et al. (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* 480(7378):490–495.
43. Pearson WR, Wood T, Zhang Z, Miller W (1997) Comparison of DNA sequences with protein sequences. *Genomics* 46(1):24–36.
44. Park C, Makova KD (2009) Coding region structural heterogeneity and turnover of transcription start sites contribute to divergence in expression between duplicate genes. *Genome Biol* 10(1):R10.
45. Rost B (1999) Twilight zone of protein sequence alignments. *Protein Eng* 12(2):85–94.
46. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797.
47. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24(8):1586–1591.
48. Gibbs RA, et al.; Rhesus Macaque Genome Sequencing and Analysis Consortium (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316(5822):222–234.
49. Makino T, McLysaght A (2010) Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci USA* 107(20):9270–9274.
50. Nazor KL, et al. (2012) Recurrent variations in DNA methylation in human pluripotent stem cells and their differentiated derivatives. *Cell Stem Cell* 10(5):620–634.
51. Wang D, et al. (2012) IMA: An R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics* 28(5):729–730.
52. Brenet F, et al. (2011) DNA methylation of the first exon is tightly linked to transcriptional silencing. *PLoS ONE* 6(1):e14524.
53. Chuang T-J, Chen F-C, Chen Y-Z (2012) Position-dependent correlations between DNA methylation and the evolutionary rates of mammalian coding exons. *Proc Natl Acad Sci USA* 109(39):15841–15846.
54. Davuluri RV, Grosse I, Zhang MQ (2001) Computational identification of promoters and first exons in the human genome. *Nat Genet* 29(4):412–417.
55. Trapnell C, et al. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* 31(1):46–53.
56. Trapnell C, et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7(3):562–578.
57. Kim SH, Yi SV (2006) Correlated asymmetry of sequence and functional divergence between duplicate proteins of Saccharomyces cerevisiae. *Mol Biol Evol* 23(5): 1068–1075.
58. Yanai I, et al. (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21(5):650–659.
59. Smit AFA, Hubley R, Green P (1996–2010) RepeatMasker Open 3.0. Available at www.repeatmasker.org.
60. Bailey TL, Bodén M, Whitington T, Machanick P (2010) The value of position-specific priors in motif discovery using MEME. *BMC Bioinformatics* 11:179.

Supporting Information

Supporting Information Corrected July 31, 2014
Keller and Yi 10.1073/pnas.1321420111

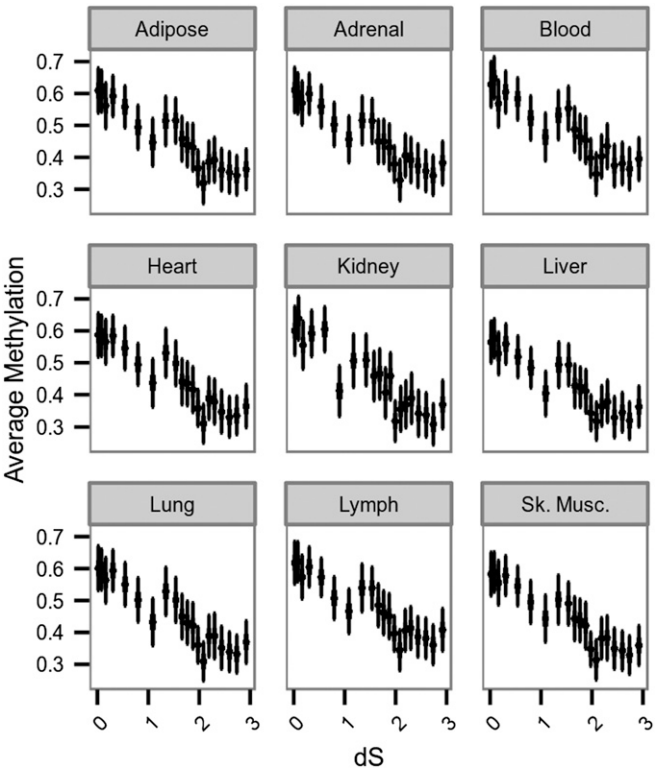

Fig. S1. Promoter methylation inversely related to evolutionary age in multiple tissues.

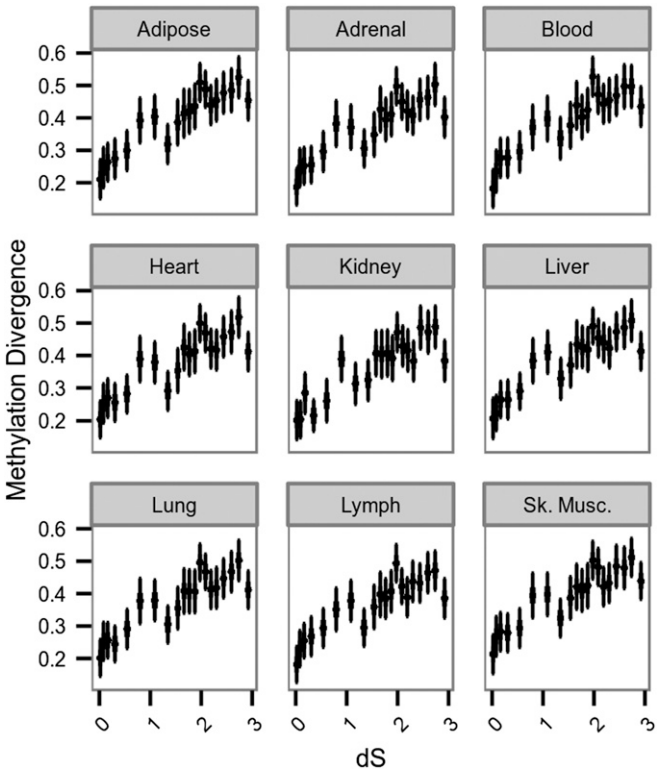

Fig. S2. Promoter methylation divergence increases with evolutionary age in multiple tissues.

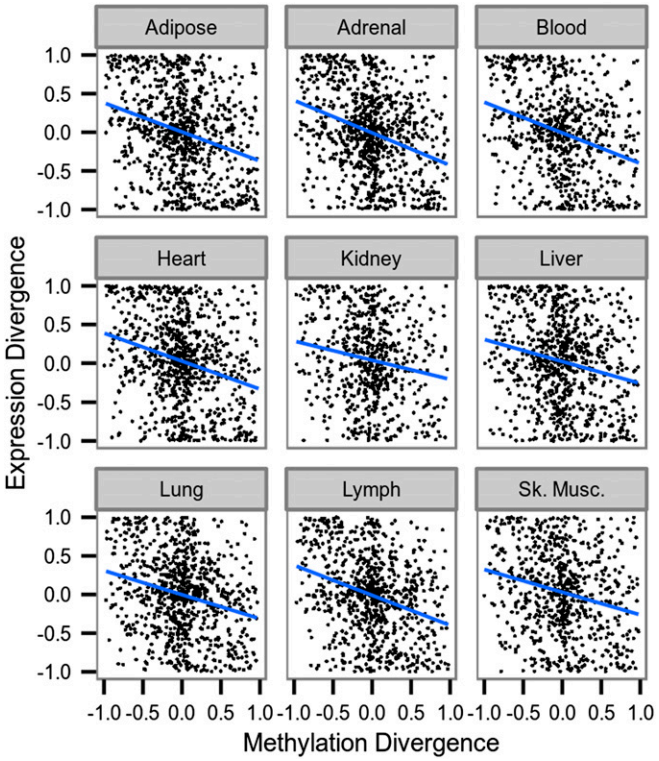

Fig. S3. Differences in methylation between duplicates are associated with differences in expression.

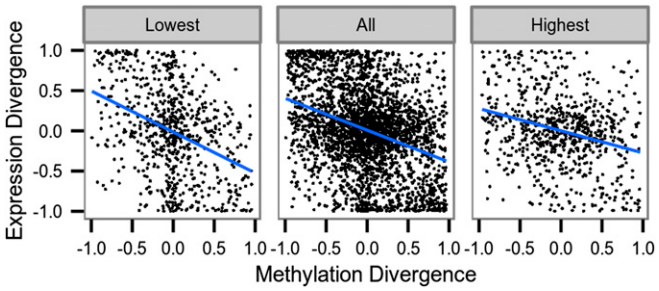

Fig. S4. Correlation between methylation and expression differences is strongest in young duplicates in brain. Duplicates were divided into four approximately equally sized bins by the number of substitutions at synonymous sites (dS). We show correlations in the first quartile (lowest, dS <2.4), all data, and the last quartile (highest, dS >7.8).

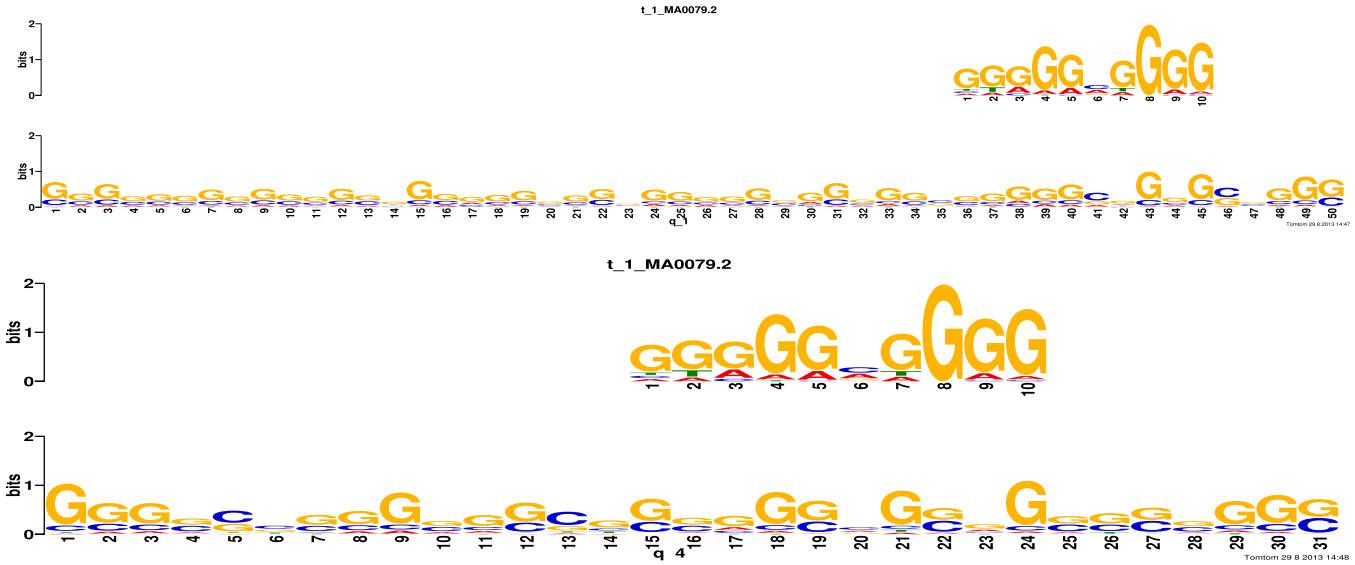

Fig. S5. Sp1 binding regions for the motifs enriched in unmethylated genes.

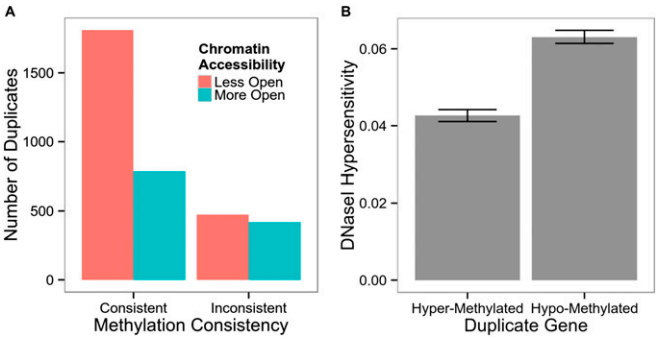

Fig. S6. Similar chromatin accessibility patterns in other ENCODE cell lines. (*A*) The frequencies of hypermethylated partners being encoded as having less (red) or more (blue) open chromatin. (*B*) The hypermethylated partners in consistently differentially methylated pairs have significantly lower DNaseI hypersensitivity compared with the hypomethylated partners.

Table S1. Relationship between DNA methylation levels of different regions and evolutionary age (measured by dS) in the whole-genome bisulfite-sequencing data

Variable 1	Variable 2	Correlation coefficient	*P* value
Promoter methylation	dS	−0.41	<2.2e-16
Gene body methylation	dS	0.002	NS (0.87)
Exon methylation	dS	0.003	NS (0.92)

Data are from prefrontal cortex DNA methylation maps in Zeng et al. (1). NS, not statistically significant.

1. Zeng J, et al. (2012) Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. *Am J Hum Genet* 91(3):455–465.

Other Supporting Information Files

Dataset S1 (TXT)
Dataset S2 (TXT)