A Simple Method for Estimating the Strength of Natural Selection on Overlapping Genes

Xinzhu Wei and Jianzhi Zhang*

Department of Ecology and Evolutionary Biology, University of Michigan

*Corresponding author: E-mail: jianzhi@umich.edu.

Accepted: December 27, 2014

Abstract

Overlapping genes, where one DNA sequence codes for two proteins with different reading frames, are not uncommon in viruses and cellular organisms. Estimating the direction and strength of natural selection acting on overlapping genes is important for understanding their functionality, origin, evolution, maintenance, and potential interaction. However, the standard methods for estimating synonymous \( d_S \) and nonsynonymous \( d_N \) nucleotide substitution rates are inapplicable here because a nucleotide change can be simultaneously synonymous and nonsynonymous when both reading frames involved are considered. We have developed a simple method that can estimate \( d_N/d_S \) and test for the action of natural selection in each relevant reading frame of the overlapping genes. Our method is an extension of the modified Nei-Gojobori method previously developed for nonoverlapping genes. We confirmed the reliability of our method using extensive computer simulation. Applying this method, we studied the longest human sense–antisense overlapping gene pair, \( LRRC8E \) and \( ENSG00000214248 \). Although \( LRRC8E \) (leucine-rich repeat containing eight family, member E) is known to regulate cell size, the function of \( ENSG00000214248 \) is unknown. Our analysis revealed purifying selection on \( ENSG00000214248 \) and suggested that it originated in the common ancestor of bony vertebrates.

Key words: synonymous substitution, nonsynonymous substitution, evolution.

Introduction

Overlapping genes generally refer to pairs of genes that overlap in their transcribed sequences. In this study, however, overlapping genes refer to pairs of genes that overlap in their protein-coding regions but use different reading frames. The first overlapping genes were discovered nearly 40 years ago in bacteriophage \( \phi \) X174 (Barrell et al. 1976). Overlapping genes have since been found in numerous viruses and cellular organisms including multicellulars such as humans, and their functional importance has been demonstrated in some case studies (Giorgi et al. 1983; Normark et al. 1983; Chen et al. 1993; Veeramachaneni et al. 2004; Pavesi 2006; Chung et al. 2008; Dornenburg et al. 2010). In theory, two genes may overlap in one of the five possible phases (fig. 1), two being sense–sense (ss) and three being sense–antisense (sas). The sas11 phase, in which the second codon position in one gene faces the third codon position in the other gene (fig. 1), was reported to be the most common type (in prokaryotes), likely because this phase minimizes the mutual constraints of the protein sequences of the overlapping genes (Rogozin et al. 2002).

To study the functionality, origin, maintenance, and evolution of overlapping genes, it is often necessary to infer the direction and strength of natural selection acting on them. The standard approach for studying natural selection acting on protein-coding genes is by estimating the ratio between the rate of nonsynonymous nucleotide substitution \( d_N \) and that of synonymous nucleotide substitution \( d_S \). However, because a mutation may be simultaneously synonymous and nonsynonymous in overlapping genes, the commonly used methods for estimating \( d_N/d_S \) are inapplicable. Several attempts have been made to estimate selection strengths in overlapping genes. Some authors treated a pair of overlapping genes as two nonoverlapping genes and calculated \( d_N/d_S \) for each gene independently using the standard methods (Yu et al. 2005; Pavesi 2006; Simon-Loriere et al. 2013). As pointed out long ago (Miyata and Yasunaga 1978), this approach is problematic, because a synonymous mutation to one of the overlapping genes may be nonsynonymous to the other gene and thus may be non-neutral. Realizing that the neutral expectation of \( d_N/d_S \) for each overlapping gene may not be 1, Nekrutenko et al. (2005) simply calculated \( d_N \)
and $d_2$ rather than their ratio, but they still applied a standard method directly to each overlapping gene. As such, the biological meanings of the estimated $d_3$ and $d_4$ are unclear. Rogozin et al. also noted the impact of one mutation on two genes and hence considered only sites that are 4-fold degenerate for one of the overlapping genes. Specifically, they were able to estimate $d_4$ for each gene in gene pairs with the ss11 phase (Rogozin et al. 2002). But this method does not apply to all overlapping genes, and estimating $d_5$ remains difficult (e.g., Rogozin et al. estimated $d_5$ from non-overlapping regions). Extending Goldman and Yang’s (1994) method for nonoverlapping coding sequences, Sabath et al. (2008) developed a maximum likelihood (ML) method for simultaneous estimation of the selection intensity in each of the two overlapping genes. However, as currently implemented, the method cannot test whether $d_4/d_5$ significantly differs from 1 for either gene (Sabath et al. 2008, 2012), rendering the utility of the method limited.

Here, we describe a simple method that estimates the selection strength of each of the two overlapping genes by separating the effects of each mutation on the two genes. Our method also estimates the associated variance, allowing a test of neutrality for each gene. We evaluate the performance of our method using computer simulation, and illustrate its utility by analyzing the human sas gene pair with the longest overlapping region.

### Materials and Methods

#### Computer Simulation

Our new method for estimating the selection strengths in overlapping genes is described in the Results section. Here, we describe the simulation used to evaluate the performance of our method. To generate a pair of overlapping genes, we set the following parameters: the overlapping phase, the length of the overlapping region $l$, the ratio ($R$) between the number of transitions and number of transversions, the distance ($d$) between two sequences defined by the expected number of substitutions per neutral site, selection strength on open reading frame 1 (ORF1) ($\omega_1$), and selection strength on ORF2 ($\omega_2$). We generated an ancestral sequence that contained overlapping ORFs by first randomly choosing sense codons for the first ORF and then removing all stop codons until no stop codon is found in each ORF. We then introduced mutations following Kimura’s (1980) two-parameter model with a preset $R$. The fixation probability of a mutation is determined jointly by $\omega_1$ and $\omega_2$. Specifically, if the mutation is synonymous in both ORFs, its fixation probability is set to be a $(0 < a < 1)$; if the mutation is synonymous to ORF1 but nonsynonymous to ORF2, its fixation probability is $a_1\omega_2$; if the mutation is synonymous to ORF2 but nonsynonymous to ORF1, its fixation probability is $a_0\omega_1$; if the mutation is nonsynonymous to both ORFs, its fixation probability is $a_0\omega_1\omega_2$. The parameter $a$ must be small enough so that $a_0\omega_1\omega_2$ are all smaller than 1. Under this scheme, both positive and negative selection can be simulated. When negative selection is simulated for both ORFs, $a$ can take any value between 0 and 1, but we assigned 0.9 to $a$ to decrease the computational time. When positive selection is simulated for ORF1 but negative selection is simulated for ORF2, 0.9/($\omega_1$ was assigned to $a$. If both ORFs are under positive selection, 0.9/($\omega_1\omega_2$) was assigned to $a$. Each ancestral sequence was evolved independently to produce two derived sequences, by either accepting or rejecting the randomly generated mutations. Simulation ended when the number of mutations introduced equals the preset number ($dl/\alpha$). $\omega_1$ and $\omega_2$ were then estimated by comparing the two simulated derived sequences. The scripts used for simulating overlapping genes and for estimating $\omega$ were written with Perl and are available at http://www.umich.edu/~zhanglab/download.htm (last accessed January 8, 2015).

#### Case Study

Annotation for human protein-coding genes and sequences used in the selection analysis were downloaded from Ensembl GRCh37 (http://useast.ensembl.org/, last accessed January 8, 2015). Overlapping genes were identified by comparing exon start and end positions of each gene on the same chromosome. For example, if exon 2 of gene A starts at position 13,780 and ends at 13,942 on Chromosome 1, and exon 5 of gene B starts at 13,950 and ends at 13,820 on the same
chromosome, we can infer that these two genes form a pair of
sas overlapping genes and that the overlapping region be-
 tween the two exons has \(13942 \text{–} 13820+1 = 123\) bp. The
overlapping genes analyzed were identified from Ensembl an-
notations using a Python script. Sequences were aligned using
an online version of ClustalW2 (http://www.ebi.ac.uk/Tools/
msa/clustalw2/, last accessed January 8, 2015). Transition/
transversion ratio was calculated using MEGA5 (Tamura
et al. 2011). The protein expression levels were from
ProteomicsDB at https://www.proteomicsdb.org/ (last
GenBank accession numbers of LRRC8 genes analyzed are
provided in supplementary tables S1 and S2, Supplementary
GenBank accession numbers of LRRC8 genes analyzed are
GenBank accession numbers of LRRC8 genes analyzed are
provided in supplementary tables S1 and S2, Supplementary
GenBank accession numbers of LRRC8 genes analyzed are
GenBank accession numbers of LRRC8 genes analyzed are
provided in supplementary tables S1 and S2, Supplementary
GenBank accession numbers of LRRC8 genes analyzed are
GenBank accession numbers of LRRC8 genes analyzed are
provided in supplementary tables S1 and S2, Supplementary
GenBank accession numbers of LRRC8 genes analyzed are
GenBank accession numbers of LRRC8 genes analyzed are
provided in supplementary tables S1 and S2, Supplementary
GenBank accession numbers of LRRC8 genes analyzed are
provided in supplementary tables S1 and S2, Supplementary
GenBank accession numbers of LRRC8 genes analyzed are
GenBank accession numbers of LRRC8 genes analyzed are
provided in supplementary tables S1 and S2, Supplementary
GenBank accession numbers of LRRC8 genes analyzed are
provided in supplementary tables S1 and S2, Supplementary

## Results

### A New Method for Estimating the Selection Strength in Overlapping Genes

Because most species use double-stranded DNA, one segment
of DNA can harbor at most six different ORFs. However, very
rarely do all six ORFs coexist. Even in cases where all six ORFs
coeexist, it is unclear whether all ORFs code for actual proteins
(Menon et al. 1990). The simplest and most common over-
lapping coding regions harbor two different ORFs, which can
be either on the same strand (ss overlap) or on opposite
strands (sas overlap) (fig. 1). The two types of ss overlap are
in fact equivalent, because they both have the third codon
positions of one ORF facing the first codon positions of the
other ORF (fig. 1). Here, we use the ss overlap as an example
to describe our method, but the same applies to all overlaps
between two ORFs.

Our method is an extension of the modified Nei-Gojobori
(nNG) method for estimating \(d_s\) and \(d_k\) in nonoverlapping
genes (Nei and Gojobori 1986; Zhang et al. 1998), but con-
siders the complication that one mutation simultaneously af-
tacts two ORFs, often with different effects. Let us consider a
pair of homologous DNA sequences (e.g., respectively from
human and mouse) that harbor overlapping ORF1 and ORF2.
Our method for quantifying the selection strength in ORF1
and that in ORF2 involves the following four steps.

In the first step, we classify human nucleotide sites in the
overlapping region into four categories depending on the im-
pacts of potential mutations on the two ORFs. The four cate-
gories are referred to as NN, NS, SN, and SS sites, respectively,
where \(N\) stands for nonsynonymous and \(S\) stands for synon-
yymous. That is, if all potential mutations at a site cause non-
synonymous changes in both ORFs, it is an NN site, and so on.
A site may belong to multiple categories and be called, for
example, 1/3 NN site and 2/3 NS site, if one-third of potential
mutations at the site cause nonsynonymous changes in both
ORFs and two-thirds of potential mutations at the site cause
nonsynonymous changes in ORF1 but synonymous changes in
ORF2. When considering potential mutations, it is important
to separate transitions from transversions because they typi-
cally have different mutation rates and have different proba-
bilities of causing nonsynonymous changes (Zhang 2000). Let
\(R\) be the ratio between the number of transitional mutations
and that of transversional mutations and be estimated from
external information (e.g., from nonoverlapping regions or
other genes). Hence, we consider a fraction of \(R/(1 + R)\) muta-
tions to be transitions and the rest transversions (Zhang et al.
1998) in determining to which of the above four categories a
site belongs. For instance, if the transitional mutation at a site
causes a synonymous change in both ORFs and the two trans-
versional mutations both cause a synonymous mutation in
ORF1 and a nonsynonymous mutation in ORF2, this site is
counted as \(R/(R + 1)\) SS site and \((1/R + 1)\) SN site. We then
calculate the total number of sites in the human overlapping
region belonging to each of the four categories. The corre-
sponding values are also calculated for the mouse sequence,
and the averaged value from the two sequences for each
category (\(L_{NN}, L_{NS}, L_{SN}, \text{ and } L_{SS}\)) will be used subsequently.

In the second step, we classify all nucleotide differences
between the two sequences into four categories: NN, NS,
SN, and SS. That is, if a difference is nonsynonymous in
both ORF1 and ORF2, it belongs to the NN group, and so
on. When a nucleotide difference is in isolation, meaning
that in neither ORF is there another difference in the same
codon as the focal difference, the classification is straight-
forward. But when a codon (in either ORF) harbors two or more
differences, the situation becomes complicated, because to
determine the categories of the multiple differences, one
has to consider all possible evolutionary pathways that can
give rise to the observed nucleotide differences. In the case
of nonoverlapping ORFs, there are two equally shortest evo-
lutionary pathways between a pair of codon sequences with
two differences (e.g., to evolve from AAA to AGG, one can go
through AAG or AGA) and six equally shortest pathways
when it harbors three differences (Nei and Gojobori 1986).
For overlapping ORFs, however, one may need to consider
a lot more pathways, because a codon in ORF1 overlaps with
a codon in ORF2, which overlaps with another codon in ORF1,
and so on. Thus, we need to find a segment of DNA in which
each codon (defined by both ORFs) has multiple nucleotide
differences with the exception of the codon at each end of
the segment (fig. 2). When this segment has a total of \(m\) nucle-
otide differences between the pair of homologous sequences,
a total of \(m!\) pathways should be considered, each of which
contains a unique order of \(m\) nucleotide changes. For each
pathway, we count the number of nucleotide changes be-
longing to each of the four categories (NN, NS, SN, and SS).
We average these numbers across all open pathways, which
are pathways with no intermediate sequences that contain
stop codons. An example is provided in supplementary
figure S1, Supplementary Material online. After classifying all

### Performance of the New Method in Estimating the Selection Strengths in Overlapping Genes

To examine the performance of the new method, we conducted extensive computer simulation of overlapping genes of each phase. The overlapping region had 3,000 nucleotides, and the simulation was repeated 100 times under each parameter set. We used exceptionally long overlapping regions to minimize the sampling error such that potential biases of our estimators became more readily detectable. We start by describing the results obtained under the ss overlap. We first examined the situation that both overlapping genes are under purifying selection. We fixed $\omega_1 = 0.2$ and $\omega_2 = 0.5$ and studied how the distance between a pair of homologous sequences affects the accuracy of estimation (fig. 3A), where the distance is defined by the expected number of substitutions per neutral site between the two homologous sequences (i.e., the expected value of $d_{SN}$). We found that the mean $\omega_1$ estimate and the mean $\omega_2$ estimate are both slightly greater than their true values, and this excess in the estimated $\omega$ value appears unrelated to the distance. This bias may be due to the fact that we simulated sequence evolution using Kimura’s two-parameter model, but estimated $d_{NN}$, $d_{NS}$, and $d_{SN}$ using the Jukes–Cantor correction, which is known to underestimate multiple hits in this scenario. When $\omega_1$ and $\omega_2$ are lower than 1, $d_{SN}$ and $d_{NS}$ are larger than $d_{NN}$, making the undercorrection more severe for the former than the latter typically have few SS sites. Thus, unless otherwise noted, we do not use $d_{SS}$ in this study. It is sometimes of interest to compare the selective pressures acting on the two overlapping genes. For this purpose, we can compute $\omega_1/\omega_2$, which equals $d_{NN}/d_{SN}$.

To calculate the variances of $d_{NN}$, $d_{NS}$, $d_{SN}$, and $d_{SS}$, the commonly used bootstrap method (Nei and Kumar 2000) is inappropriate because of the difficulty in bootstrapping codons from one ORF while maintaining the other ORF. We therefore extend an approximate analytical method previously developed for estimating the variances of $d_{NN}$ and $d_{NS}$ in the Nei–Gojobori method (Nei 1987), which is known to be quite accurate (Ota and Nei 1994). Following this method, we calculate the variance of $d_{NN}$ by $\text{Var}(d_{NN}) = \text{Var}(d_{NN})/((1-4d_{NN}/3)^2$, where the variance of $p_{NN}$ is given by $\text{Var}(p_{NN}) = p_{NN}(1-p_{NN})/L_{NN}$. Variances of $d_{NS}$, $d_{SN}$, and $d_{SS}$ can be similarly estimated. Standard deviations (SDs) of $d_{NN}$, $d_{NS}$, $d_{SN}$, and $d_{SS}$ are then estimated by taking the square root of their variances, respectively. The hypothesis of neutral evolution of ORF1 can be tested by a Z-test of the equality between $d_{NN}$ and $d_{SN}$. That is, we can conduct a Z-test using $Z = (d_{NN} - d_{SN})/(\text{Var}(d_{NN}) + \text{Var}(d_{SN}))^{1/2}$. Similarly, the neutral evolution hypothesis for ORF2 can be tested by a Z-test of the equality between $d_{NN}$ and $d_{NS}$. We can also test if the strengths of natural selection acting on the two ORFs are equal by a Z-test of the equality between $d_{SN}$ and $d_{NS}$.

---

**Table 1:** Determining the shortest overlapping region for mutational pathway consideration. Shown is an example of the ss overlap. Codons in ORF1 are marked with lines above the sequences, whereas codons in ORF2 are marked with lines below the sequences. Differences between the two species are in black, whereas identical nucleotides are in grey. The boxed region is the shortest region for mutational pathway consideration.

<table>
<thead>
<tr>
<th>Species</th>
<th>Codons in ORF1</th>
<th>Codons in ORF2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>ATG GCC AGC</td>
<td>ATG GCC GC</td>
</tr>
<tr>
<td>Mouse</td>
<td>ATG GCC AC</td>
<td>ATG GCC GC</td>
</tr>
</tbody>
</table>

---

**Fig. 2.** Determining the shortest overlapping region for mutational pathway consideration. Shown is an example of the ss overlap. Codons in ORF1 are marked with lines above the sequences, whereas codons in ORF2 are marked with lines below the sequences. Differences between the two species are in black, whereas identical nucleotides are in grey. The boxed region is the shortest region for mutational pathway consideration.
and the resultant $\omega_1$ and $\omega_2$ upward biased. Nevertheless, the biases appear to be generally lower than 10%. In contrast, if we estimate $\omega_1$ and $\omega_2$ by the mNG method without considering the mutual influences between overlapping genes, the estimates are much higher than their respective true values (fig. 3A). This is because some synonymous mutations to one ORF are nonsynonymous to the other ORF and hence have been removed by purifying selection, causing overestimation of $\omega_1$ and $\omega_2$. Because the true $\omega_1 < \omega_2 < 1$, $\omega_2$ is overestimated to a larger extent than $\omega_1$ (fig. 3A).

Next, we examined the situation that one overlapping gene is under positive selection ($\omega_1 = 3$), while the other is under purifying selection ($\omega_2 = 0.2$). We again found the mean estimates of $\omega_1$ and $\omega_2$ by our method to be close to their respective true values, for all levels of distance considered (fig. 3B). When the mNG method is used, $\omega_2$ is slightly underestimated (fig. 3B), likely because some synonymous mutations to ORF2 are beneficial to ORF1 and are fixed by positive selection. In contrast, $\omega_1$ is grossly overestimated by mNG (fig. 3B), for the reason mentioned in the previous paragraph.

We next examined the impact of the transition/transversion ratio $R$ on estimates of $\omega_1$ and $\omega_2$ when their true values are 0.2 and 1, respectively (fig. 3C). We found both $\omega_1$ and $\omega_2$ slightly overestimated. This becomes moderately severe for $\omega_2$ when $R \geq 10$, probably due to the aforementioned undercorrection of multiple hits by the Jukes–Cantor formula that is more serious when $R$ gets higher. The mNG method performs similarly well as the new method in estimating $\omega_1$ (fig. 3C), likely because of the lack of any selection on ORF2. But $\omega_2$ is grossly overestimated by mNG (fig. 3C). Because ORF2 itself is not under any selection, the above phenomenon must be due to the fact that synonymous mutations to ORF2 are more likely than nonsynonymous mutations to ORF2 to be deleterious to ORF1.

We next varied $\omega_1$ from 0.2 to 3.0 while keeping $\omega_2$ at 0.2. We found estimates of $\omega_1$ and $\omega_2$ by our method to be generally reliable (fig. 3D). By contrast, $\omega_1$ is consistently and grossly overestimated by mNG, whereas $\omega_2$ is underestimated when $\omega_1 < 1$ and underestimated when $\omega_1 > 1$, as expected (fig. 3D).
In addition to the ss overlap, we also examined the three sas overlapping phases with different parameter sets. We found that our method generated reliable results under all phases (fig. 4). In contrast, the mNG method can make grossly wrong estimates, and the direction and extent of the error depend on \( o_1 \), \( o_2 \), and the specific overlapping phase (fig. 4). For phase sas12, third codon positions in ORF1 overlap with third codon positions in ORF2. Consequently, the fraction of SS sites is higher than that in other phases, allowing the possibility of estimating natural selection using SS sites. We thus also estimated \( o_1 \) by \( d_{NS}/d_{SS} \) and estimated \( o_2 \) by \( d_{SN}/d_{SS} \) for phase sas12 (see sas12* in fig. 4). The results showed that these estimates are either similar to or slightly better than those using NN sites (see sas12 in fig. 4).

Because the analytical formulas for SDs are approximate, we used computer simulation to investigate their accuracies. For the ss phase, we examined the reliabilities of the analytically computed SD(\( d_{NN} \)), SD(\( d_{NS} \)), and SD(\( d_{SN} \)), but could not examine SD(\( d_{SS} \)) because of the paucity of SS sites. We conducted 100 simulation replications under each set of parameters. We then compared the SD among the 100 \( d_{rel} \) values obtained and the mean of SD(\( d_{NN} \)) analytically calculated using the data from each simulation. The same was done for \( d_{NS} \) and \( d_{SN} \). We found the analytically calculated SD values to be overall similar to the simulation observations, with statistically insignificant differences (fig. 5).

### Evolutionary Analysis of the Human Gene Pair with the Longest Sense–Antisense Overlapping Region

To illustrate the utility of our method, we searched for an appropriate pair of overlapping genes from Ensembl for detailed analysis. We found that Ensembl annotates most ss overlapping genes with different reading frames as alternative splicing (Curwen et al. 2004), greatly underestimating the prevalence of ss overlapping genes. We thus focused on sas overlapping and identified the longest sas overlapping coding region in the human genome, containing 732 bases. The involved genes are \( LRRC8E \) (leucine-rich repeat containing eight family, member E) and an uncharacterized gene with an Ensembl Gene ID of \( ENSG00000214248 \). The structure of this gene pair (fig. 6A) shows that the entire 243 amino acid

---

**Fig. 4.**—Performances of the new (NEW) method and modified Nei–Gojobori (mNG) method in estimating the selection intensities (\( o_1 \) and \( o_2 \)) on simulated overlapping genes of various phases indicated on the x axis. Each symbol represents the mean from 100 replications under a given parameter set, and error bars show the standard error. In each panel, the parameters are listed above the panel, whereas different overlapping phases are shown on the x axis. The results for sas12* are estimates using SS sites (i.e., \( o_1 = d_{NS}/d_{SS} \) and \( o_2 = d_{SN}/d_{SS} \)) under the sas12 phase.
coding region of ENSG00000214248 lies within the second exon of LRRC8E, with the sas12 overlapping phase. It was recently discovered that LRRC8E functions as an essential component of the cell volume-regulated anion channel VRAC (Voss et al. 2014), but whether ENSG00000214248 encodes a functional protein and what its function is are unknown.

We found from the recently published human proteomic data (Wilhelm et al. 2014) that ENSG00000214248 is not only transcribed but also translated in coronary sinus and blood platelet (Fig. 6B). The protein expression sites of ENSG00000214248 and those of LRRC8E overlap in blood platelet but are otherwise distinct (fig. 6B). The expression levels of the two replicates are generally comparable (fig. 6B).

We acquired the sequences of the orthologous genes of human ENSG00000214248 and LRRC8E from the macaque genome sequence. Using our method, we estimated the \( \omega \) values for the two genes in the overlapping region as well as the \( \omega \) in the nonoverlapping region of LRRC8E. \( R \) was estimated to be 3.61 from the nonoverlapping region of LRRC8E using Kimura’s (1980) two-parameter model. We found that the overlapping region and the nonoverlapping region of LRRC8E have been under similar levels of purifying selection, with \( \omega = 0.08 \) and 0.09, respectively. The \( \omega \) for ENSG00000214248 is 0.20, significantly lower than the neutral expectation of 1 (\( P < 0.002 \), two-tail Z-test), suggesting that this uncharacterized gene has been under purifying selection at least since the divergence between human and macaque. For the overlapping region, we used SS sites in the above estimation of the \( \omega \) values for ENSG00000214248, because there was no substitution at NS sites.

Because ENSG00000214248 is entirely within LRRC8E, we traced the origin of ENSG00000214248 by examining its presence in LRRC8E of various species. We were able to identify LRRC8E in all bony vertebrate genome sequences available at Ensembl and NCBI, but not in shark, lamprey, or any invertebrate genome. Interestingly, we also identified the ORF of ENSG00000214248 within LRRC8E in most bony vertebrates, including zebrafish (fig. 6C). Apparently, ENSG00000214248 already existed in the common ancestor of bony vertebrates, but was pseudogenized several times in subsequent evolution (fig. 6C). Because LRRC8E is a member of the LRRC8 family that contains five genes in human, we reconstructed the phylogeny of this gene family (supplementary fig. S2, Supplementary Material online) to investigate if ENSG00000214248 originated before LRRC8E. We discovered that the closest relative to LRRC8E is LRRC8C, which can be found in bony vertebrates and shark.
presumable ENSG00000214248 reading frame in LRRC8C contains several premature stop codons in each species examined (human, macaque, mouse, rat, zebrafish, and shark), suggesting that the common ancestor of LRRC8C and LRRC8E did not contain ENSG00000214248. Thus, the antisense reading frame probably originated in LRRC8E shortly after the birth of LRRC8E from the duplication of LRRC8C.

**Discussion**

Overlapping genes have been identified in many species and are particularly common in bacteria and viruses (Normark et al. 1983; Veeramachaneni et al. 2004), but their evolutionary studies have been hampered by the inapplicability of the standard methods for inferring natural selection acting on overlapping genes. We developed a simple method to estimate the selection strength on each of the overlapping ORFs and demonstrated the reliability of our method by computer simulation. Our method allows testing whether an overlapping gene is under natural selection and hence can be used to identify functional genes from hypothetical overlapping reading frames, as was demonstrated in the example of ENSG00000214248.

To more readily detect potential biases of our method, we simulated long overlapping regions (3,000 sites). In reality, however, overlapping regions are much shorter. We also
performed simulations using overlapping regions of 750 sites and 300 sites, respectively (supplementary fig. S3, Supplementary Material online), based on the parameters used in figure 3A and B. When the overlapping region is short and the distance is low, many sequences had no substitution in NS sites or SN sites, making our method inapplicable. For cases where our method did work, the mean $\omega$ estimates were reasonably good, although the standard errors were large, as expected (supplementary fig. S3, Supplementary Material online). Thus, accurately estimating $\omega$ values of short overlapping regions remains challenging unless the divergence between the two taxa compared is high. On the basis of current annotations of eukaryotic genomes, there are not many overlapping genes that have long evolutionary histories. However, as in the example studied, although the orthologs of human ENSG00000214248 are present in many vertebrates, they have not been annotated outside primates. It is likely that much more overlapping genes and long-lasting overlapping genes than currently annotated exist. Overlapping genes are prevalent in viral genomes. Many viruses have high mutation rates, allowing the use of our methods even for relatively short overlapping regions.

Sabath et al. (2008) noted that the ML method they developed does not perform well under low distances (mean sequence divergence across sites $< 8\%$). To examine if our method suffers from the same problem, we compared the two methods using the parameters in figure 3A and B. The results showed that the two methods are similar in their sensitivity to distance (supplementary fig. S4, Supplementary Material online). However, under both negative (supplementary fig. S4a, Supplementary Material online) and positive (supplementary fig. S4b, Supplementary Material online) selection, our method outperforms the ML method in terms of the accuracy of the $\omega$ estimates.

Although we introduced our method in the context of estimating the selective strength using interspecific comparisons, our method may also be applied to intraspecific data or comparisons between intraspecific and interspecific data. For instance, let us use $D_{NN}$, $D_{NS}$, $D_{SN}$, and $D_{SS}$ to denote the numbers of the four types of substitutions in a pair of overlapping genes, respectively, and use $p_{NN}$, $p_{NS}$, $p_{SN}$, and $p_{SS}$ to denote the corresponding numbers of the four types of polymorphisms, respectively. We can conduct a selection test similar to the McDonald–Kreitman test (McDonald and Kreitman 1991) for ORF1 by comparing $D_{NN}$, $D_{NS}$, $p_{NN}$, and $p_{SN}$, because $D_{NN}$ equals $D_{SN}$ under the null hypothesis of neutrality. Similarly, we can test selection in ORF2 by comparing $D_{NN}$, $D_{NS}$, $p_{NN}$, and $p_{SN}$. In addition to studying overlapping genes, our method can also be applied to the study of the functionality of certain alternative splicing. Alternative splicing is generally demonstrated by the existence of various transcripts from a gene, but the existence of a transcript is not a proof that the transcript is functional. For splice variants using alternative reading frames, our method may be used to test if the alternative reading frame has been under natural selection, which would support the functionality of the splice variant.

In summary, we believe that our development of a simple method for estimating the selective strengths on overlapping genes will facilitate researches toward understanding the origin, evolution, and functionality of overlapping genes.

**Supplementary Material**

Supplementary figures S1–S4 and tables S1–S2 are available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org/).

**Acknowledgments**

We thank Wei-Chin Ho and Jian-Rong Yang for valuable comments. This work was supported in part by the US National Institutes of Health research grant R01GM103232 to J.Z.

**Literature Cited**


