J Mol Evol (2011) 73:305–315
DOI 10.1007/s00239-011-9477-9

Is There a Twelfth Protein-Coding Gene in the Genome of Influenza A? A Selection-Based Approach to the Detection of Overlapping Genes in Closely Related Sequences

Niv Sabath · Jeffrey S. Morris · Dan Graur

Received: 30 March 2011 / Accepted: 2 December 2011 / Published online: 21 December 2011

Abstract Protein-coding genes often contain long overlapping open-reading frames (ORFs), which may or may not be functional. Current methods that utilize the signature of purifying selection to detect functional overlapping genes are limited to the analysis of sequences from divergent species, thus rendering them inapplicable to genes found only in closely related sequences. Here, we present a method for the detection of selection signatures on overlapping reading frames by using closely related sequences, and apply the method to several known overlapping genes, and to an overlapping ORF on the negative strand of segment 8 of influenza A virus (NEG8), for which the suggestion has been made that it is functional. We find no evidence that NEG8 is under selection, suggesting that the intact reading frame might be non-functional, although we cannot fully exclude the possibility that the method is not sensitive enough to detect the signature of selection acting on this gene. We present the limitations of the method using known overlapping genes and suggest several approaches to improve it in future studies. Finally, we examine alternative explanations for the sequence conservation of NEG8 in the absence of selection. We show that overlap type and genomic context affect the conservation of intact overlapping ORFs and should therefore be considered in any attempt of estimating the signature of selection in overlapping genes.

Keywords Overlapping genes · Influenza A · Gene discovery

Background

Discovering a new gene in a clinically important virus is an exciting proposition, because gene number is usually very small and each additional gene greatly increases the proteome repertoire and, hence, the list of potential targets for pharmaceutical intervention. For example, the eleventh protein-coding gene in the influenza A genome, *PB1-F2* (which overlaps *PB1*), was discovered 20 years after the initial annotation of its genome, thereby increasing the proteome repertoire by 10% (Chen et al. 2001). Recently, it was suggested that the negative strand of segment 8 in influenza A codes for an additional protein (Clifford et al. 2009; Zhirnov et al. 2007). This open-reading frame (ORF) (Fig. 1), which was noted when this segment was first sequenced (Baez et al. 1980), is intact in several human influenza A viruses, but disrupted in non-human influenza A viruses, such as avian viruses, as well as in influenza B

Electronic supplementary material The online version of this article (doi:10.1007/s00239-011-9477-9) contains supplementary material, which is available to authorized users.

N. Sabath (✉)
Institute of Evolutionary Biology and Environmental Studies, University of Zurich, 8057 Zurich, Switzerland
e-mail: nsabath@gmail.com

N. Sabath
The Swiss Institute of Bioinformatics, Basel, Switzerland

J. S. Morris
Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX 77230, USA
e-mail: jefmorris@mdanderson.org

D. Graur
Department of Biology and Biochemistry, University of Houston, Houston, TX 77204, USA
e-mail: dgraur@uh.edu

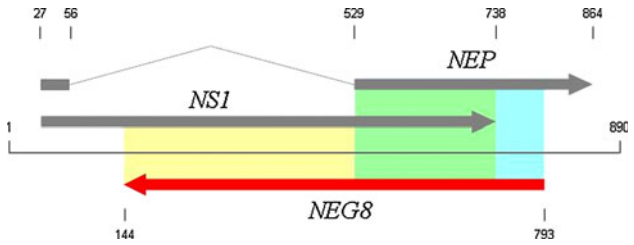

Fig. 1 Schematic representation of segment 8 in human influenza A viruses. The location of the hypothetical gene, *NEG8*, is shown in *red*. The *NS1–NEG8* overlap is marked in *yellow*. The *NS1–NEG8–NEP* triple overlap is marked in *green*. The short *NEG8–NEP* overlap is marked in *blue* (Color figure online)

and C viruses. Two main indications that this hypothetical gene, called *NEG8*, may be functional were given: (1) The ORF has been conserved in human influenza A viruses for almost a century (Clifford et al. 2009; Zhirnov et al. 2007) and (2) an epitope (a short peptide) encoded by this ORF was reported to induce an immune system response through cytotoxic T cells isolated from mice infected with this virus (Clifford et al. 2009; Zhong et al. 2003).

We note, however, that it is fairly common for at least one of the five possible overlapping reading frames of any gene to contain an ORF of a length that may be suitable to encode a protein, but it is extremely difficult to ascertain whether such an intact overlapping ORF is functional or not. Methods for prediction of protein-coding genes search for (1) the presence of an intact ORF, (2) evidence of mRNA expression, and (3) evolutionary conservation. In the case of overlapping genes, however, these properties are often uninformative because: (1) intact overlapping ORFs that are nonetheless non-functional are expected to be fairly common, (2) both same-strand and opposite-strand overlapping ORFs may be transcribed regardless of functionality (Lavorgna et al. 2004), and (3) non-functional overlapping ORFs are evolutionary conserved because of their sharing a sequence with functional genes. In addition, viral overlapping genes could originate de novo (Keese and Gibbs 1992), in which case a lack of conservation is expected in a recent origin of the gene. The situation in the literature is quite confused; on the one hand, many overlapping ORFs have been deemed upon reexamination as functional (Chung et al. 2008; Firth 2008; Firth and Atkins 2008a; b, 2009; Sabath et al. 2009), and on the other, the functionality of numerous annotated overlapping genes have been questioned (Palleja et al. 2008; Sabath and Graur 2010; Silke 1997; Williams et al. 2009).

Ultimately, whether an ORF is functional or not could only be determined experimentally. However, the rapid accumulation of sequence data calls for improvement of computational methods for prediction of functional overlapping genes. The commonest way to computationally predict functional overlapping genes is to identify purifying selection, which is tightly associated with functionality. However, the identification of selection signature in overlapping genes is complicated by the sequence interdependence between two overlapping coding regions (Miyata and Yasunaga 1978; Smith and Waterman 1981), which vary among overlap types (Krakauer 2000). Several attempts at estimating selection intensity in overlapping genes reported inordinate degrees of positive selection (e.g., Campitelli et al. 2006; Hughes et al. 2001; Li et al. 2004; Obenauer et al. 2006). In some studies, one gene was found to exhibit positive selection while the overlapping gene showed signs of strong purifying selection (e.g., Campitelli et al. 2006; Hughes et al. 2001; Li et al. 2004; Obenauer et al. 2006). Inferences of positive selection in overlapping genes have been questioned (Holmes et al. 2006; Pavesi 2007; Sabath et al. 2008b; Suzuki 2006), mostly because ignoring overlap constraints might bias selection estimates. Rogozin et al. (2002) tried to overcome this problem by focusing on sites in which all changes are synonymous in one gene and nonsynonymous in the overlapping gene. This method, however, is only practical when dealing with one type of overlap.

A model for the nucleotide substitutions in overlapping genes was introduced by Hein and Stovlbaek (1995), who followed approximate models for non-overlapping genes that classify sites according to degeneracy classes (Li et al. 1985; Nei and Gojobori 1986; Pamilo and Bianchi 1993). This model was later incorporated into a method for annotation of viral genomes (de Groot et al. 2007; McCauley et al. 2007; McCauley and Hein 2006), and recently used for estimating selection on overlapping genes (de Groot et al. 2008). Pedersen and Jensen (2001) suggested a non-stationary substitution model for overlapping reading frames that extended the codon-based model of Goldman and Yang (1994). This model encompasses the evolutionary process more accurately than the approximate model (Hein and Stovlbaek 1995) by accounting for position dependency of each site in an overlap region (Pedersen and Jensen 2001). However, this improvement disallowed the straightforward estimation of parameters and forced the authors to apply a computationally expensive simulation procedure (Pedersen and Jensen 2001). Firth and Brown (2005) proposed a method, suitable for sequence pairs, that calculates several statistics for each particular pairwise sequence alignment and uses a Monte Carlo simulation to determine whether the sequence is single-coding or double-coding. This method led to the discovery of novel overlapping genes in many viral taxa (Chung et al. 2008; Firth 2008; Firth and Atkins 2008a, b, 2009). Sabath et al. (2008b) devised a new method within a maximum-likelihood framework to fit a non-stationary Markov model of codon substitution to data from two aligned orthologous overlapping sequences. By using this method, Sabath et al.

(2009) have predicted the existence of a new overlapping gene in the genomes of four viruses, including the Israeli acute paralysis virus, which is implicated in the colony-collapse disorder of honeybees. This prediction was later supported by Firth et al. (2009). A comparison between the methods of Firth and Brown and Sabath et al. (Sabath and Graur 2010) showed differences in performances across overlap types with Sabath et al.'s method exhibiting higher sensitivity on average. Other methods for detecting overlapping genes, which do not use selection as a criterion, have been proposed in the literature (Chung et al. 2007; Nekrutenko and He 2006; Nekrutenko et al. 2005; Neuhaus et al. 2010; Ribrioux et al. 2008; Szklarczyk et al. 2007; Trifonov and Rabadan 2009; Xu et al. 2010). These methods vary widely in their logic, efficacy, and applicability.

Unfortunately, all of the above methods are unsuitable for dealing with sequences exhibiting high levels of similarity (e.g., Firth and Brown 2005, Fig. 5; Sabath et al. 2008b, Fig. 4), such as data from influenza subtypes (Bao et al. 2008). Further, these two methods were designed to detect the signature of selection on pairs of sequences without accounting for the phylogenetic relationships among multiple sequences. Firth and Brown (2006) have attempted to overcome this issue by only using neighboring terminal pairs of taxa. Although, this approach enables a uniform sampling of the tree branches, it also has the downside of including highly similar pairs of sequences, which (as noted above) yield inaccurate inferences.

Here, we propose a new method for the detection of purifying selection on hypothetical overlapping reading frames. The method infers evolutionary changes along a phylogenetic tree of closely related sequences using maximum-parsimony criteria. Two studies have used a similar approach to detect site-specific positive selection in non-overlapping genes. Fitch et al. (1997) constructed a maximum-parsimony tree of the *HA*1 gene of influenza A, inferred all changes along the branches, and tested if the changes are randomly distributed among the positions of the sequence. Similarly, Suzuki and Gojobori (1999) used the neighbor-joining method (Saitou and Nei 1987) to reconstruct a phylogenetic tree and adapted the pairwise method of Nei and Gojobori (1986) to estimate positive selection on a single site in human leukocyte antigen gene, HIV-1 *ENV* gene, and the influenza A *HA1* gene. The results of both studies show that even a weak signature of selection, as in the case of site-specific positive selection, could be detected in closely related sequences. Our method makes use of purifying selection, which unlike positive selection, is known to affect the vast majority of sites in protein-coding genes and can, therefore, be used to detect functional overlapping genes.

Materials and Methods

Rationale

To detect the signature of purifying selection acting on a hypothetical gene, we employed the principle that nonsynonymous mutations are generally more deleterious than synonymous mutations. If a hypothetical gene is under purifying selection, a mutation that is nonsynonymous in both genes is expected to be more deleterious than one that is nonsynonymous in one gene and synonymous in the other.

Algorithm

Substitutions are classified into transitions and transversions. Substitutions in the known gene (k) that overlaps a hypothetical gene (h) are further classified into four categories: nonsynonymous in both genes (N_kN_h), nonsynonymous in the known gene and synonymous in the hypothetical gene (N_kS_h), synonymous in the known gene and nonsynonymous in the hypothetical gene (S_kN_h), and synonymous in both genes (S_kS_h). Taken together, we define eight categories of substitutions for each pair of overlapping sequences (Table 1).

Throughout this article, we use the term "category pair" to denote a pair of substitutional categories that differ only in the hypothetical gene (i.e., N_kN_h vs. N_kS_h, and S_kN_h vs. S_kS_h). The four category pairs are set apart in bold cells in Table 1. For example, being nonsynonymous versus synonymous in the hypothetical gene is the only difference between a transitional mutation in the N_kN_h category versus a transitional mutation in the N_kS_h category. In the absence of selection on the hypothetical gene, the rates of the two substitutional categories within a pair should be equal to each other. If, on the other hand, the hypothetical gene is functional and under purifying selection, the rate of

Table 1 Notation of the test variables and specification of category pairs

	Categories			
	N_kN_h	N_kS_h	S_kN_h	S_kS_h
Transitions				
Possible substitutions	$\boldsymbol{P_1}$	$\boldsymbol{P_2}$	P_3	P_4
Observed substitutions	$\boldsymbol{O_1}$	$\boldsymbol{O_2}$	O_3	O_4
Expected	$\boldsymbol{E_1}$	$\boldsymbol{E_2}$	E_3	E_4
Transversions				
Possible substitutions	$\boldsymbol{P_5}$	$\boldsymbol{P_6}$	P_7	P_8
Observed substitutions	$\boldsymbol{O_5}$	$\boldsymbol{O_6}$	O_7	O_8
Expected	$\boldsymbol{E_5}$	$\boldsymbol{E_6}$	E_7	E_8

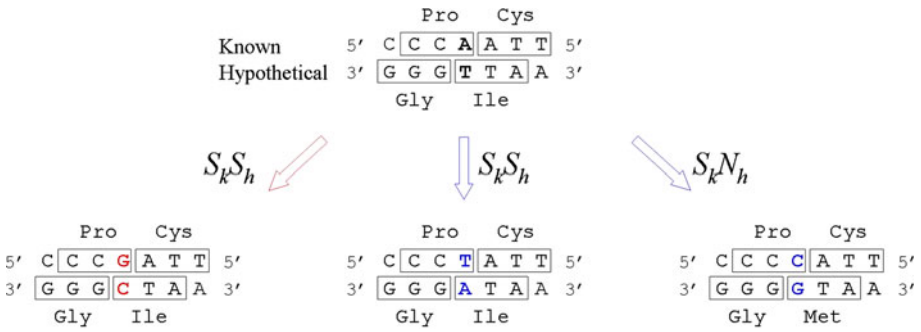

Fig. 2 Three possible substitutions at site 4 (in *bold*) in the phase-2 opposite-strand overlapping sequence between a known gene and a hypothetical gene. Transitions and transversions are marked in *red* and *blue*, respectively. The substitutional category of each change is noted (Color figure online)

substitution in the N_kN_h category should be lower than that in N_kS_h, because a nonsynonymous change in both the known and the hypothetical genes will affect two gene products rather than one. Similarly, the rate of change in the S_kN_h category should be lower than that in S_kS_h.

For example, Fig. 2 illustrates three possible substitutions at site 4 in a phase-2 opposite-strand overlapping sequence. If the hypothetical gene is under selection, the change A/T → C/G (S_kN_h category) is expected to be more deleterious than the change A/T → T/A (S_kS_h category), which does not affect the amino acid sequence of the hypothetical gene. We note that there is no assumption about the intensity of selection on the known gene and, hence, the method can be used even when the known gene is in fact under no selection (evolving under strict neutrality).

Given a multiple alignment of closely related DNA sequences, the method includes four steps:

(1) Construction of an unrooted phylogenetic tree.
(2) Reconstruction of ancestral sequences.
(3) Classification of the changes along the tree into the eight substitutional categories.
(4) Testing for the signature of purifying selection through comparisons between category pairs that differ in the hypothetical gene only.

We used PAUP (Swofford 2003) to construct an unrooted neighbor-joining tree (Saitou and Nei 1987) of each data set and assigned the ancestral character states of the internal nodes using the parsimony criteria (Fitch 1971). Using the reconstructed sequences, we counted the number of unique observed substitutions (O) in each category along all the branches. We used the unique number of substitutions rather than the total number of substitutions to minimize the possible biases from non-uniform sampling (e.g., in industrial countries where more isolates are collected) and from highly constrained variable sites, in which only a few character states are permissible (Delport et al. 2008). For any given sequence of length n, there are $3n$ possible substitutions (e.g., Fig. 2) that can be classified into these eight categories (Table 1). For any given set, we calculated the number of possible substitutions (P) in each category as the average across the sequences in all nodes of the tree.

We use the ratio O_i/P_i as a measure of the rate of substitutions in category i. If the hypothetical gene is not under selection, we expect no difference between O_i/P_i and O_j/P_j where i and j are two categories that differ only in the hypothetical gene: $\langle i, j\rangle \in \{\langle 1, 2\rangle, \langle 3, 4\rangle, \langle 5, 6\rangle, \langle 7, 8\rangle\}$.

The null hypothesis of no selection on the hypothetical gene is defined as:

$$\frac{O_i}{P_i} = \frac{O_j}{P_j} = \frac{(O_i + O_j)}{(P_i + P_j)}. \quad (1)$$

Under this null hypothesis, we estimate the expected values of O_i and O_j to be:

$$E_i = P_i \frac{(O_i + O_j)}{(P_i + P_j)} \text{ and } E_j = P_j \frac{(O_i + O_j)}{(P_i + P_j)}. \quad (2)$$

We, then, construct a contingency table for each category pair

$$\begin{pmatrix} O_i & O_j \\ E_i & E_j \end{pmatrix}, \langle i, j\rangle \in \{\langle 1, 2\rangle, \langle 3, 4\rangle, \langle 5, 6\rangle, \langle 7, 8\rangle\}. \quad (3)$$

This contingency table is used to test the null hypothesis. For example, O_1 and O_2, which are the observed number of substitutions in the transitional N_kN_h and N_kS_h categories (Table 1), differ only by being nonsynonymous or synonymous in the hypothetical gene. E_1 and E_2, which are the expected values of O_1 and O_2, are estimated based on the null hypothesis in which the rate of substitutions in the two categories is equal. If the hypothetical gene is subjected to selection, any change in the N_kN_h category would affect both genes and O_1 is expected to be lower than E_1, whereas O_2 is expected to be higher than E_2.

We used the one-tailed Fisher's exact test (1925) to determine significance of the negative association, in which the observations tend to lie in the lower left and upper right of the table. For example, in the contingency table $\begin{pmatrix} O_1 & O_2 \\ E_1 & E_2 \end{pmatrix}$, the alternative, in which category 1 is under stronger purifying selection, requires that O_1 is lower than E_1, and that O_2 is higher than E_2. Because, the test requires exact numbers, the expected values were rounded. Finally,

we combined the four P values into a single test statistic using Fisher's method (1925).

In the case of overlap types other than phase-2 opposite-strand overlap (the overlap type of *NEG8*), the number of possible S_kS_h substitutions is very small (Table 2). This makes the S_kS_h categories, and consequently the N_kS_h–S_kS_h pair, uninformative. Therefore, we focused on the two category pairs of N_kN_h and N_kS_h (bolded in Table 1) in the test.

Sequence Data

There are several viruses, which have been extensively sequenced and are, hence, suitable for analysis by our method. We compiled sequences of (1) the most sequenced subtypes of influenza A (H3N2, H1N1, and H5N1), (2) influenza B, (3) human immunodeficiency virus 1 (HIV-1), (4) human papillomavirus type 16 (HPV-16), and (5) hepatitis B virus (HBV). The data consists of three different kinds of sets: (1) *NS1–NEG8* overlaps in influenza A, H3N2 and H1N1 subtypes, in which the *NEG8* ORF is intact, (2) *NS1–NEG8* overlaps in influenza A, H5N1 subtype and influenza B, in which the *NEG8* ORF is disrupted, and (3) seven known same-strand overlapping genes: *PB1–PB1-F2* overlaps from influenza A, H3N2 subtype; *PB1–PB1-F2* overlaps and *NS1–NEP* overlaps from influenza A, H5N1 subtype; *NA-NB* overlaps from influenza B; *ENV-REV* overlaps from HIV-1; *E2–E4* overlaps from HPV-16; and *large-S-polymerase(large-S-POL)* overlaps from HBV. The four *NS1–NEG8* sets are listed in Table 3. The seven sets of same-strand overlapping genes are listed in Table 4.

For each set of sequences of overlapping genes from influenza viruses, we obtained multiple alignments of all full-length sequences excluding sequences with insertions and/or deletions from the NCBI Influenza Virus Resource (Bao et al. 2008). Because, ancestral sequence reconstruction is inaccurate for divergent sequences (Zhang and Nei 1997), we also excluded sequences from early isolates (before 1990) that result in very long braches on the tree. For the *NEG8* sets, we only analyzed the region of *NEG8* that overlaps with *NS1* (382 bases), and excluded regions of triple overlap (*NS1–NEG8–NEP*) and the short region of *NEP–NEG8* overlap (Fig. 1). The nucleotide sequences of *ENV–REV* overlap from HIV-1, *E2–E4* overlap from HPV-16, and *large-S-polymerase (POL)* overlap from HBV, were obtained from the complete genomes of these viruses (Accessions NC_001802, NC_001526, and NC_003977, respectively). For each of the above three genes, we used nucleotide Blast (Altschul et al. 1990) to find all full-length sequences excluding sequences with insertions and/or deletions. For all data sets, the frequencies of the possible and observed numbers of substitutions in each category are listed in Table 2.

To gather a better understanding of the new method and the data, we used our previous method (SLG; Sabath et al. 2008b), which is applicable to pairs of sequences. For each data set, we chose 100 random sequence pairs with 5% divergence or higher (when available). For each pair, we applied the method twice: first, to estimate the intensity of selection on both genes simultaneously (Sabath et al. 2008b), and second, using the likelihood ratio test for two hierarchical models (Sabath et al. 2009). In model 1, we assume no selection on the overlapping gene. In model 2, the overlapping gene is assumed to be under selection. If model 2 fits the data significantly better than model 1 ($P < 0.01$), then the overlapping gene is predicted to be under selection and is most probably functional.

We note, however, that the methods are not fully comparable for four reasons: (1) SLG is not suitable for short evolutionary distances and therefore we excluded sequence pairs whose divergence from each other is <5%. (2) Because, SLG is only applicable for pairs of sequences, and the current data sets contain a huge number of pairs (e.g., the HBV set contains more than 700,000 pairs), we had to restrict the computation to only a small fraction of the data (100 pairs). (3) The randomly chosen pairs cannot be treated as independent data points because of their evolutionary relationships; hence, there is no simple solution for multiple-testing correction. (4) The choice of pairs leads to internal branches on the tree to be over-represented in the analysis.

Finally, we used the complete genomes of 768 RNA non-ambisense viruses (i.e., viruses that utilize only one strand to code for proteins) to evaluate the influence of genome composition on the probability of having an overlapping ORF. Genomes were obtained from NCBI. Stop codon frequencies in the five possible reading frames [on the same strand in phases 1 and 2, and on the opposite strand in phases 0, 1, and 2] were calculated from the coding sequences of each genome.

The source code (written in Matlab) and the data files used in this study can be accessed at http://overlappinggene.sourceforge.net.

Results

Our data consist of three sets: (1) *NS1* genes with disrupted *NEG8* ORFs, (2) *NS1* genes with intact *NEG8* ORFs, and (3) known same-strand overlapping genes. Because, our method is indifferent to the degree of selection acting on the known gene, we applied it on all sets twice (reciprocally) as a control. For example, for the sets in which no *NEG8* ORF exists (H5N1 and influenza B), we tested for selection on the *NS1* genes to examine the method's ability to detect true genes and also tested for selection on the

Table 2 Possible and observed number of substitutions in each category

	Gene 1	Gene 2	Ts/Tv		Nonsynonymous substitutions in gene 2		Synonymous substitutions in gene 2	
					NN	*NS*	*SN*	*SS*
Influenza A: H3N2	*NEG8*	*NS1*	Ts	*P*	237.8	16.9	22.9	104.4
				O	73	10	16	68
			Tv	*P*	554.9	78.6	65.8	64.8
				O	30	10	16	7
Influenza A: H1N1	*NEG8*	*NS1*	Ts	*P*	237.8	18.0	18.4	107.8
				O	53	4	12	61
			Tv	*P*	547.1	81.9	70.3	64.7
				O	25	4	13	7
Influenza A: H5N1	*NEG8*	*NS1*	Ts	*P*	237.2	19.8	17.0	108.1
				O	111	12	14	93
			Tv	*P*	539.7	84.4	71.0	69.0
				O	61	13	13	21
Influenza B	*NEG8*	*NS1*	Ts	*P*	346.5	32.3	21.3	167.8
				O	104	11	12	114
			Tv	*P*	829.9	141.1	92.4	72.7
				O	66	12	18	10
Influenza A: H3N2	*PB1-F2*	*PB1*	Ts	*P*	97.9	83.1	86.0	3.0
				O	21	13	76	3
			Tv	*P*	377.0	73.0	89.0	1.0
				O	16	4	19	1
Influenza A: H5N1	*PB1-F2*	*PB1*	Ts	*P*	97.9	83.1	86.0	3.0
				O	15	15	81	4
			Tv	*P*	382.4	70.5	86.0	1.0
				O	20	4	23	1
Influenza A: H5N1	*NS1*	*NEP*	Ts	*P*	62.6	53.1	55.2	0.1
				O	29	25	36	0
			Tv	*P*	236.8	57.9	44.2	3.0
				O	22	7	9	0
Influenza B	*NB*	*NA*	Ts	*P*	92.6	93.3	103.2	2.0
				O	29	33	62	1
			Tv	*P*	358.0	99.8	121.4	2.9
				O	25	7	13	0
HIV-1	*ENV*	*REV*	Ts	*P*	72.2	76.1	72.8	3.9
				O	14	29	23	3
			Tv	*P*	296.2	65.3	86.5	1.9
				O	27	12	12	0
HPV-16	*E2*	*E4*	Ts	*P*	87.0	97.0	94.0	4.0
				O	3	13	1	1
			Tv	*P*	352.8	112.8	97.4	1.1
				O	4	5	2	0
HBV	*Pol*	*Large S*	Ts	*P*	408.4	378.0	385.8	20.8
				O	183	264	220	24
			Tv	*P*	1471.1	457.7	442.5	14.6
				O	401	204	149	14

Ts transitions; *Tv* transversions

Table 3 Sets of *NEG8–NS1* overlaps

	Virus	Number of sequences	Gene 1	*P*	Gene 2	*P*	ω_1 (STD)	ω_2 (STD)	Gene 1% Pos[a]	Gene 2% Pos[a]
Hypothetical *NEG8* gene	Influenza A: H3N2	410	*NEG8*	0.151	*NS1*	**0.017**	NA	NA	NA	NA
	Influenza A: H1N1	217	*NEG8*	0.667	*NS1*	**0.008**	0.978 (0.157)	0.305 (0.052)	0	91
No *NEG8* gene	Influenza A: H5N1	581	*NEG8*	0.359	*NS1*	0.086	0.817 (0.184)	0.470 (0.112)	0	18
	Influenza B	229	*NEG8*	0.604	*NS1*	**0.015**	0.734 (0.155)	0.487 (0.081)	0	10

[a] Percentage of positives out of 100 random pairs

Table 4 Sets of known same-strand overlapping genes

Virus	Number of sequences	Gene 1	*P*	Gene 2	*P*	ω_1 (STD)	ω_2 (STD)	Gene 1% Pos[a]	Gene 2% Pos[a]
Influenza A: H3N2	999	*PB1-F2*	0.446	*PB1*	$\mathbf{2.0 \times 10^{-7}}$	0.864 (0.120)	0.045 (0.011)	0	100
Influenza A: H5N1	522	*PB1-F2*	0.690	*PB1*	$\mathbf{5.6 \times 10^{-10}}$	0.640 (0.204)	0.024 (0.013)	34	100
Influenza A: H5N1	581	*NS1*	0.647	*NEP*	0.151	0.808 (0.250)	0.505 (0.168)	6	33
Influenza B	165	*NB*	0.617	*NA*	0.044	0.847 (0.303)	0.394 (0.147)	0	44
HIV-1	323	*REV*	0.265	*ENV*	0.072	2.373 (1.295)	1.370 (1.028)	0	1
HPV-16	63	*E4*	0.606	*E2*	0.062	NA	NA	NA	NA
HBV	561	*Large S*	**0.008**	*Pol*	$\mathbf{1.1 \times 10^{-7}}$	1.783 (0.749)	0.422 (0.085)	26	97

[a] Percentage of positives out of 100 random pairs

disrupted *NEG8* ORF to find whether the method yields false positive results. We distinguish between the results for the *NEG8–NS1* sets (Table 3) and those for same-strand overlaps (Table 4), because in the later case, the type of overlap dictates using the reduced version of the method (see "Materials and Methods") and is therefore predicted to have a lower power. In both tables, *P* values lower then 0.05 are bolded. For reference, we used SLG, our previously developed method (Sabath et al. 2008b) to evaluate the intensity of selection on each gene and the percentage of pairs, which were predicted to be under purifying selection out of the total number of pairs (as in Sabath and Graur 2010; Sabath et al. 2009).

NEG8–NS1 Sets

We found significant signatures of selection in three out of the four known *NS1* genes (the *P* value of the fourth is relatively low, 0.086), demonstrating the ability of the method to detect selection in known functional genes (Table 3). We used the two sets in which the *NEG8* ORF is disrupted (H5N1 and influenza B) to verify that the method does not yield false positive inferences. In both cases, no signature of selection was identified on *NEG8*. Finally, we applied the method to test for selection on the hypothetical *NEG8* ORF in the H1N1 and H3N2 sets. We did not find a significant signature of selection on the *NEG8* ORF in either case.

Known Same-Strand Overlapping Genes

We used seven sets of known overlapping genes in five viruses to test the performance of the method in same-strand overlapping genes (Table 4). Because, there are very few possible substitutions that are synonymous in both genes (Table 2), the test is applied only to two category pairs. We find significant signatures of selection on both genes in the *large-S-POL* overlap set, demonstrating the ability of the method to detect selection on same-strand overlaps. In three sets (*PB1–PB1-F2* from H3N2 and H5N1 subtypes, and *NA–NB*), we identified significant signatures of selection on one gene while no selection was identified on the other. In two sets (*ENV–REV* and *E2–E4*), the *P* value of one gene is relatively low (0.072 and 0.062, respectively) while no selection was identified on the other. For the last set (*NS1–NEP*), there were no significant signatures of selection detected on any of the two genes.

Comparison to SLG

Although, SLG (Sabath et al. 2008b) is not fully comparable to our present method, it complements the new method by providing estimates of selection intensity and the percentage of positive pairs, i.e., pairs that were predicted to be under purifying selection. Similar to the new method, SLG did not detect any signature of selection in the two sets in which the *NEG8* ORF is disrupted. There

are four cases in which SLG predicted more than 90% positive pairs, all were also identified by the new method. In four other cases, there is a moderate signal (26–44% positive pairs). For two of these cases (*NEP* and *PB1–PB1-F2* from H5N1 subtype) we did not detect selection by the new method. In two cases, we found a weak signal (10 and 18%); one was detected by the new method, the other was not. Finally, in six cases we found 0–6% positive pairs.

Discussion

We present a new method for the detection of functional overlapping genes utilizing the signature of selection for closely related sequences. The method detects selection signatures by extending the principle that nonsynonymous mutations are generally more deleterious than synonymous mutations. In overlapping genes, this principle is translated into the following expectation: a mutation that is nonsynonymous in both genes is expected to be more deleterious than a mutation that is nonsynonymous in one gene and synonymous in the other.

The specificity of the method was evaluated by using influenza viruses in which the *NEG8* ORF is disrupted. Evaluating the sensitivity of the method, in contrast, is a much more difficult proposition. Ideally, we would have liked to use known opposite-strand overlapping genes with phase-2 overlaps as in *NEG8*. Unfortunately, there is no sufficient data for all the currently known opposite-strand overlapping genes (e.g., Todd et al. 2001). We, therefore, chose to use known same-strand overlapping genes, in which the sensitivity is expected to be lower because of the fewer category pairs that can be used. Nevertheless, the method has detected the signature of selection on both genes in one set and one gene in three other sets.

Variation in selection pressures among sites may also affect the method's performance. For example, a mutation of the N_kS_h category at a constrained site of the known protein may be more deleterious than a mutation of the N_kN_h category at less constrained sites of both genes. As a control for site variation, we used data sets of orthologous sequences that share constrained sites, but in which the hypothetical *NEG8* ORF is disrupted. In future studies, it may be beneficial to incorporate information of known constrained sites into the model. In addition to variation in selection pressures among sites within one gene, difference in the intensities of selection acting on the two overlapping genes may also affect the performance of the method, especially when the hypothetical gene is under considerable weaker purifying selection than the known overlapping gene. In an overlapping gene pair, the newer gene is expected to be under weaker purifying selection (Liang and Landweber 2006), because it has evolved for a shorter period of time than its overlapping genes as well as under the constraints of its overlapping gene. The hypothetical overlapping gene would usually be the newer gene. Therefore, detection of new overlapping genes by signature of purifying selection is difficult. Indeed, *PB1-F2*, a novel human influenza A gene (Chen et al. 2001) was not detected by the method (Table 4). This gene was shown to be under selective pressure that is weaker by an order of magnitude than that on the older overlapping gene, *PB1* (Sabath et al. 2008b). The lack of selection signature in the *NEG8* ORF may be due to the intact reading frame being non-functional. However, we cannot fully exclude the possibility that the method cannot detect the signature of selection because the gene is too new.

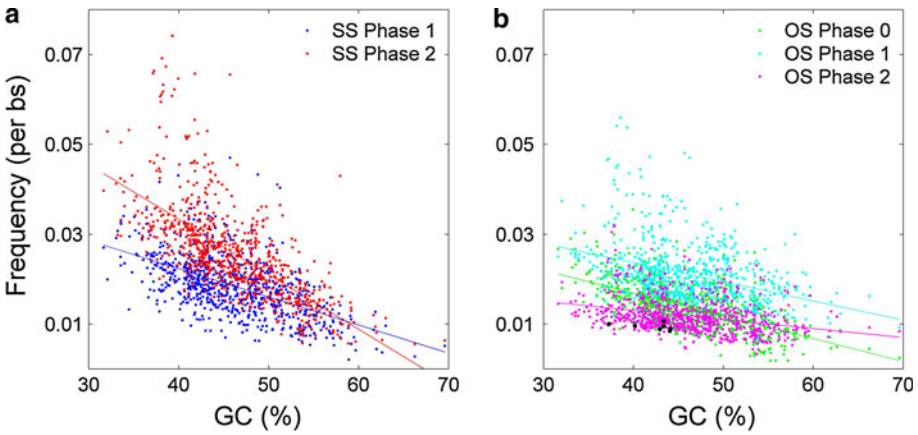

Fig. 3 Frequencies of stop codons of 768 RNA viruses in the five possible reading frames plotted against genomic GC content. **a** Stop codon frequencies on the same strand (SS) in phase 1 (*blue*) and phase 2 (*red*). **b** Stop codon frequencies on the opposite strand (OS) in phase 0 (*green*), phase 1 (*cyan*), and phase 2 (*magenta*). Stop codons frequencies on the opposite strand in phase 2 (*NEG8* phase) of influenza genomes are marked in *black* (Color figure online)

We consider two factors that may contribute to the conservation of intact overlapping ORFs in the absence of selection. In a non-functional ORF that overlaps a functional gene, the potential for a stop codon mutation that disrupts the ORF is determined by the sequence of the functional gene. Consequently, the frequencies of stop codons vary among the five overlap types (Fig. 3), leading to higher number of non-functional overlapping ORFs in overlap types with low frequency of stop codons (Sabath et al. 2008a; Silke 1997). Indeed, the specific overlap type between ~90% of the *NEG8* ORF and *NS1* (phase-2 opposite-strand overlap) was found to have the lowest frequency of stop codons from among all genes (Fig. 3). Moreover, influenza genomes have a below-average frequency of stop codons in this phase (Fig. 3b, black dots) increasing the probability of non-functional ORFs. This observation joins previous studies, which have also studied the frequency of stop codons with relation to overlapping genes (e.g., Nekrutenko et al. 2005; Ribrioux et al. 2008; Trifonov and Rabadan 2009). Finally, the triple overlap involving *NEG8*, *NS1*, and *NEP* may increase the conservation of *NEG8* because any change in this region is likely to be nonsynonymous in either *NS1* or *NEP* or both. These two factors indicate that conservation of an overlapping ORF is strongly dependent on its genomic context. Hence, it is important to account for genomic context in any future attempt to detect overlapping genes.

The method presented here belongs to a group of approximate methods, in which sites are classified by degeneracy class. Approximate methods are useful for analyses of large data sets, but are less accurate than maximum-likelihood methods (Yang and Nielsen 2000). Hence, more powerful methods may be developed within the maximum-likelihood framework. An additional parameter that could be incorporated in future methods is the time of origin of each substitution, which could be estimated by using the sampling dates as calibration (Drummond and Rambaut 2007). Deleterious substitutions are expected to be more prevalent among new substitutions (Pybus et al. 2007) because they had a lower chance of being eliminated from the population by selection. Therefore, it would be beneficial to account for the age of each nucleotide substitution when selection is estimated at the population level. The comparison of the new method to a previous one (Sabath et al. 2008b) revealed little difference in performance. Finally, the inability of the method to detect several known overlapping genes requires careful interpretation of negative results. However, since none of the existing methods in the literature is applicable to this type of data, we believe that our method constitutes an important contribution and would be helpful for any future study aimed at detecting selection signatures in overlapping genes.

Availability and Requirements

Project name: detection of overlapping genes
Project home page: http://overlappinggene.sourceforge.net
Operating system(s): Platform independent
Programming language: Matlab
Other requirements: Matlab statistics, optimization, and bioinformatics toolboxes
License: GPL
Any restrictions to use by non-academics: license needed

Acknowledgments We thank Dr. Chris Upton for suggesting this problem and providing valuable information. DG and NS were supported by a Small Grant Award from the University of Houston and by the US National Library of Medicine grant LM010009-01 to DG and Giddy Landan.

Conflict of interest The authors declare that they have no conflict of interest.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Baez M, Taussig R, Zazra JJ, Young JF, Palese P, Reisfeld A, Skalka AM (1980) Complete nucleotide sequence of the influenza A/PR/8/34 virus NS gene and comparison with the NS genes of the A/Udorn/72 and A/FPV/Rostock/34 strains. Nucleic Acids Res 8:5845–5858

Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, Lipman D (2008) The influenza virus resource at the National Center for Biotechnology Information. J Virol 82:596–601

Campitelli L, Ciccozzi M, Salemi M, Taglia F, Boros S, Donatelli I, Rezza G (2006) H5N1 influenza virus evolution: a comparison of different epidemics in birds and humans (1997–2004). J Gen Virol 87:955–960

Chen W, Calvo PA, Malide D, Gibbs J, Schubert U, Bacik I, Basta S, O'Neill R, Schickli J, Palese P, Henklein P, Bennink JR, Yewdell JW (2001) A novel influenza A virus mitochondrial protein that induces cell death. Nat Med 7:1306–1312

Chung WY, Wadhawan S, Szklarczyk R, Pond SK, Nekrutenko A (2007) A first look at ARFome: dual-coding genes in mammalian genomes. PLoS Comput Biol 3:e91

Chung BY, Miller WA, Atkins JF, Firth AE (2008) An overlapping essential gene in the Potyviridae. Proc Natl Acad Sci USA 105:5897–5902

Clifford M, Twigg J, Upton C (2009) Evidence for a novel gene associated with human influenza A viruses. Virol J 6:198

de Groot S, Mailund T, Hein J (2007) Comparative annotation of viral genomes with non-conserved gene structure. Bioinformatics 23:1080–1089

de Groot S, Mailund T, Lunter G, Hein J (2008) Investigating selection on viruses: a statistical alignment approach. BMC Bioinform 9:304

Delport W, Scheffler K, Seoighe C (2008) Frequent toggling between alternative amino acids is driven by selection in HIV-1. PLoS Pathog 4:e1000242

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7:214
Firth AE (2008) Bioinformatic analysis suggests that the Orbivirus VP6 cistron encodes an overlapping gene. Virol J 5:48
Firth AE, Atkins JF (2008a) Bioinformatic analysis suggests that a conserved ORF in the waikaviruses encodes an overlapping gene. Arch Virol 153:1379–1383
Firth AE, Atkins JF (2008b) Bioinformatic analysis suggests that the Cypovirus 1 major core protein cistron harbours an overlapping gene. Virol J 5:62
Firth AE, Atkins JF (2009) Analysis of the coding potential of the partially overlapping 3′ ORF in segment 5 of the plant fijiviruses. Virol J 6:32
Firth AE, Brown CM (2005) Detecting overlapping coding sequences with pairwise alignments. Bioinformatics 21:282–292
Firth AE, Brown CM (2006) Detecting overlapping coding sequences in virus genomes. BMC Bioinform 7:75
Firth AE, Wang QS, Jan E, Atkins JF (2009) Bioinformatic evidence for a stem-loop structure 5′-adjacent to the IGR-IRES and for an overlapping gene in the bee paralysis dicistroviruses. Virol J 6:193
Fisher R (1925) Statistical methods for research workers. Oliver and Boyd, Edinburgh
Fitch WM (1971) Toward defining the course of evolution: minimum change for a specified tree topology. Syst Zool 20:406–416
Fitch WM, Bush RM, Bender CA, Cox NJ (1997) Long term trends in the evolution of H(3) HA1 human influenza type A. Proc Natl Acad Sci USA 94:7712–7718
Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736
Hein J, Stovlbaek J (1995) A maximum-likelihood approach to analyzing nonoverlapping and overlapping reading frames. J Mol Evol 40:181–189
Holmes EC, Lipman DJ, Zamarin D, Yewdell JW (2006) Comment on "Large-scale sequence analysis of avian influenza isolates". Science 313:1573 author reply 1573
Hughes AL, Westover K, da Silva J, O'Connor DH, Watkins DI (2001) Simultaneous positive and purifying selection on overlapping reading frames of the tat and vpr genes of simian immunodeficiency virus. J Virol 75:7966–7972
Keese PK, Gibbs A (1992) Origins of genes: "big bang" or continuous creation? Proc Natl Acad Sci USA 89:9489–9493
Krakauer DC (2000) Stability and evolution of overlapping genes. Evol Int J Org Evol 54:731–739
Lavorgna G, Dahary D, Lehner B, Sorek R, Sanderson CM, Casari G (2004) In search of antisense. Trends Biochem Sci 29:88–94
Li WH, Wu CI, Luo CC (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150–174
Li KS, Guan Y, Wang J, Smith GJ, Xu KM, Duan L, Rahardjo AP, Puthavathana P, Buranathai C, Nguyen TD, Estoepangestie AT, Chaisingh A, Auewarakul P, Long HT, Hanh NT, Webby RJ, Poon LL, Chen H, Shortridge KF, Yuen KY, Webster RG, Peiris JS (2004) Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. Nature 430:209–213
Liang H, Landweber LF (2006) A genome-wide study of dual coding regions in human alternatively spliced genes. Genome Res 16:190–196
McCauley S, Hein J (2006) Using hidden Markov models and observed evolution to annotate viral genomes. Bioinformatics 22:1308–1316
McCauley S, de Groot S, Mailund T, Hein J (2007) Annotation of selection strengths in viral genomes. Bioinformatics 23:2978–2986
Miyata T, Yasunaga T (1978) Evolution of overlapping genes. Nature 272:532–535
Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426
Nekrutenko A, He J (2006) Functionality of unspliced XBP1 is required to explain evolution of overlapping reading frames. Trends Genet 22:645–648
Nekrutenko A, Wadhawan S, Goetting-Minesky P, Makova KD (2005) Oscillating evolution of a mammalian locus with overlapping reading frames: an XLalphas/ALEX relay. PLoS Genet 1:e18
Neuhaus K, Oelke D, Fürst D, Scherer S, Keim DA (2010) Towards automatic detecting of overlapping genes—clustered BLAST analysis of viral genomes. In: Proceedings of the 8th European conference on evolutionary computation, machine learning and data mining in bioinformatics (EvoBIO '10)
Obenauer JC, Denson J, Mehta PK, Su X, Mukatira S, Finkelstein DB, Xu X, Wang J, Ma J, Fan Y, Rakestraw KM, Webster RG, Hoffmann E, Krauss S, Zheng J, Zhang Z, Naeve CW (2006) Large-scale sequence analysis of avian influenza isolates. Science 311:1576–1580
Palleja A, Harrington ED, Bork P (2008) Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? BMC Genomics 9:335
Pamilo P, Bianchi NO (1993) Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Mol Biol Evol 10:271–281
Pavesi A (2007) Pattern of nucleotide substitution in the overlapping nonstructural genes of influenza A virus and implication for the genetic diversity of the H5N1 subtype. Gene 402:28–34
Pedersen AM, Jensen JL (2001) A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. Mol Biol Evol 18:763–776
Pybus OG, Rambaut A, Belshaw R, Freckleton RP, Drummond AJ, Holmes EC (2007) Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. Mol Biol Evol 24:845–852
Ribrioux S, Brungger A, Baumgarten B, Seuwen K, John MR (2008) Bioinformatics prediction of overlapping frameshifted translation products in mammalian transcripts. BMC Genomics 9:122
Rogozin IB, Spiridonov AN, Sorokin AV, Wolf YI, Jordan IK, Tatusov RL, Koonin EV (2002) Purifying and directional selection in overlapping prokaryotic genes. Trends Genet 18:228–232
Sabath N, Graur D (2010) Detection of functional overlapping genes: simulation and case studies. J Mol Evol 71:308–316
Sabath N, Graur D, Landan G (2008a) Same-strand overlapping genes in bacteria: compositional determinants of phase bias. Biol Direct 3:36
Sabath N, Landan G, Graur D (2008b) A method for the simultaneous estimation of selection intensities in overlapping genes. PLoS ONE 3:e3996
Sabath N, Price N, Graur D (2009) A potentially novel overlapping gene in the genomes of Israeli acute paralysis virus and its relatives. Virol J 6:144
Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425
Silke J (1997) The majority of long non-stop reading frames on the antisense strand can be explained by biased codon usage. Gene 194:143–155
Smith TF, Waterman MS (1981) Overlapping genes and information theory. J Theor Biol 91:379–380
Suzuki Y (2006) Natural selection on the influenza virus genome. Mol Biol Evol 23:1902–1911

Suzuki Y, Gojobori T (1999) A method for detecting positive selection at single amino acid sites. Mol Biol Evol 16:1315–1328

Swofford DL (2003) PAUP*. Phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, MA

Szklarczyk R, Heringa J, Pond SK, Nekrutenko A (2007) Rapid asymmetric evolution of a dual-coding tumor suppressor INK4a/ARF locus contradicts its function. Proc Natl Acad Sci USA 104:12807–12812

Todd D, Weston JH, Soike D, Smyth JA (2001) Genome sequence determinations and analyses of novel Circoviruses from Goose and Pigeon. Virology 286:354–362

Trifonov V, Rabadan R (2009) The contribution of the PB1-F2 protein to the fitness of Influenza A viruses and its recent evolution in the 2009 Influenza A (H1N1) pandemic virus. PLoS Curr 1:RRN1006

Williams TA, Wolfe KH, Fares MA (2009) No rosetta stone for a sense-antisense origin of aminoacyl tRNA synthetase classes. Mol Biol Evol 26:445–450

Xu H, Wang P, Fu Y, Zheng Y, Tang Q, Si L, You J, Zhang Z, Zhu Y, Zhou L, Wei Z, Lin B, Hu L, Kong X (2010) Length of the ORF, position of the first AUG and the Kozak motif are important factors in potential dual-coding transcripts. Cell Res 20:445–457

Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17:32–43

Zhang J, Nei M (1997) Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. J Mol Evol 44(Suppl 1):S139–S146

Zhirnov OP, Poyarkov SV, Vorob'eva IV, Safonova OA, Malyshev NA, Klenk HD (2007) Segment NS of influenza A virus contains an additional gene NSP in positive-sense orientation. Dokl Biochem Biophys 414:127–133

Zhong W, Reche PA, Lai CC, Reinhold B, Reinherz EL (2003) Genome-wide characterization of a viral cytotoxic T lymphocyte epitope repertoire. J Biol Chem 278:45135–45144