J Mol Evol (2010) 71:308–316
DOI 10.1007/s00239-010-9386-3

Detection of Functional Overlapping Genes: Simulation and Case Studies

Niv Sabath · Dan Graur

Received: 18 December 2009 / Accepted: 26 July 2010 / Published online: 4 September 2010

Abstract As far as protein-coding genes are concerned, there is a non-zero probability that at least one of the five possible overlapping sequences of any gene will contain an open-reading frame (ORF) of a length that may be suitable for coding a functional protein. It is, however, very difficult to determine whether or not such an ORF is functional. Recently, we proposed a method that predicts functionality of an overlapping ORF if it can be shown that it has been subject to purifying selection during its evolution. Here, we use simulation to test this method under several conditions and compare it with the method of Firth and Brown. We found that under most conditions, our method detects functional overlapping genes with higher sensitivity than Firth and Brown's method, while maintaining high specificity. Further, we tested the hypothesis that the two aminoacyl tRNA synthetase classes have originated from a pair of overlapping genes. A central piece of evidence ostensibly supporting this hypothesis is the assertion that an overlapping ORF of a heat-shock protein-70 gene, which exhibits some similarity to class 2 aminoacyl tRNA synthetases, is functional. We found signature of purifying selection only in highly divergent sequences, suggesting that the method yields false-positives in high sequence divergence and that the overlapping ORF is not a functional gene. Finally, we examined three cases of overlap in the human genome. We find varying signatures of purifying selection acting on these overlaps, raising the possibility that two of the overlapping genes may not be functional.

Keywords Overlapping genes · Purifying selection · Annotation

Introduction

Methods for prediction of protein-coding genes use three properties: (1) the presence of open-reading frame (ORF), (2) expression of mRNA, and (3) evolutionary conservation. These properties, however, are often uninformative in the case of overlapping genes 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 their sequence is shared with functional genes. As a result, annotation programs often fail to correctly predict functional overlapping genes (Delcher et al. 1999).

Much work has been done in attempts to distinguish functional from spurious overlapping genes. Silke (1997) showed that the frequency of opposite-strand overlapping ORFs in vertebrate genomes is highly influenced by genomic GC content and codon usage, suggesting that a large proportion of these ORFs may be spurious. Nekrutenko and colleagues (Chung et al. 2007; Nekrutenko and He 2006; Nekrutenko et al. 2005; Szklarczyk et al. 2007) investigated overlapping genes in the human genome

Electronic supplementary material The online version of this article (doi:10.1007/s00239-010-9386-3) contains supplementary material, which is available to authorized users.

An abstract of this paper accompanied a poster presentation at the 13th Annual International Conference on Computational Molecular Biology (RECOMB) in May 2009.

N. Sabath (✉) · D. Graur
Department of Biology and Biochemistry,
University of Houston, Houston, TX 77204, USA
e-mail: nsabath@gmail.com

and introduced a set of methods for functionality prediction based on several overlap properties including overlap length, the probability of substitution to a stop codon, and mutation patterns. Liang and Landweber (2006) predicted that ~7% of human genes contain overlapping alternatively spliced exons, based on similarity to known proteins at the levels of sequence, and secondary and tertiary structures, as well as the existence of motifs known to exist in functional genes. Ribrioux et al. (2008) scanned alternative reading frames in orthologous human–mouse–rat genes for amino-acid conservation, length, repeats, and the probability of mutations to a stop codon to predict several new candidate overlaps. Palleja et al. (2008) examined the conservation of length between overlapping genes in different bacterial species and concluded that many of the long overlapping genes have been misannotated. Xu et al. (2010) combined experimental and computational analyses to explore how ORF length, the position of the first AUG codon, and the Kozak motif affect translation of overlapping transcripts.

Although all of the above studies have used the underlying principle that genomic features of gene overlap would be conserved by purifying selection when functional, none has estimated selection pressure directly. The reason is that estimation of selection intensity is complicated by the sequence interdependence between the two genes (Miyata and Yasunaga 1978). Firth and Brown (2005) were the first to use direct estimations of selection to detect functional overlapping genes. Their method (FB), which is suitable for sequence pairs, 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 was later applied to multiple sequences by choosing neighboring terminal pairs of taxa in a phylogenetic tree (Firth and Brown 2006). With the FB method, novel overlapping genes were discovered in many viral taxa (Chung et al. 2008; Firth 2008; Firth and Atkins 2008a, b, 2009). We developed a different method (SLG) for the estimation of selection intensities in overlapping genes (Sabath et al. 2008b). SLG uses a maximum-likelihood framework to fit a Markov model of codon substitution to data from two aligned orthologous overlapping sequences. Later, we used SLG to detect signature of purifying selection in unannotated overlapping ORFs (Sabath et al. 2009). In order to detect a signature of purifying selection, we estimate the likelihood of two hierarchical models. In model 1, there is no selection on the ORF. In model 2, the ORF is assumed to be under selection. The likelihood-ratio test is used to test whether model 2 fits the data significantly better than model 1, in which case, the ORF is predicted to be under selection and most probably functional. With this method, we predicted the existence of a new functional overlapping gene in the genomes of four viruses within the Dicistroviridae family (Sabath et al. 2009). This prediction was later supported by Firth et al. (2009). Other methods that make use of selection signatures to detect functional overlapping genes were proposed (de Groot et al. 2007, 2008; McCauley et al. 2007), but these methods have seldom been used, most probably due to the lack of accessible implementation.

Here, we use simulation to test and compare the FB and SLG methods under several conditions. We discuss the influence of overlap type, selection pressure, sequence divergence, and sequence length on the performance of the methods. We also examine in detail a purported case of overlap between the gene encoding heat-shock protein-70 (*HSP70*) and an opposite-strand ORF (*OS-ORF*) (Carter and Duax 2002; Konstantopoulou et al. 1995; Monnerjahn et al. 2000; Rother et al. 1997; Silke 1997). This overlap was deemed the "Rosetta stone" for the origin of the two aminoacyl tRNA synthetase (aaRS) classes from opposite-strand overlapping genes, based on a perceived similarity between *OS-ORF* and class 1 aaRS, on the one hand, and between *HSP70* and class 2 aaRS, on the other hand (Carter and Duax 2002; Rodin and Ohno 1995) (Fig. 1). Recently,

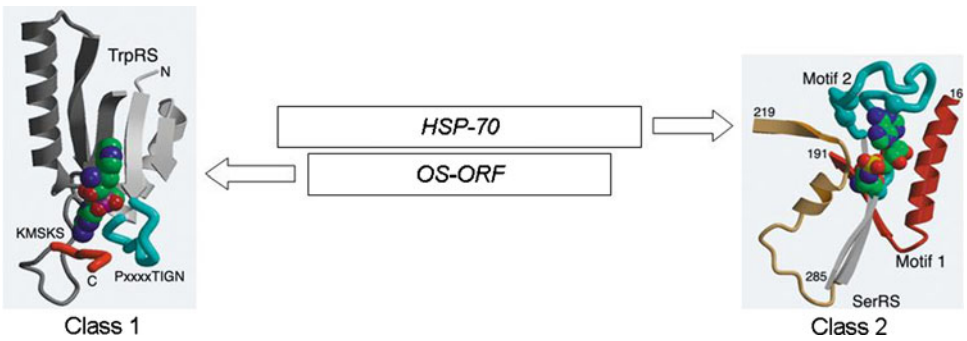

Fig. 1 The "Rosetta stone" hypothesis. The two aminoacyl tRNA synthetase (aaRS) classes were proposed to originate from opposite-strand overlapping genes, based on a perceived similarity between *OS-ORF* and class 1 aaRS, on the one hand, and between *HSP70* and class 2 aaRS, on the other hand (Carter and Duax 2002; Rodin and Ohno 1995, images from Carter and Duax 2002)

the functionality of *OS-ORF* was questioned by Williams et al. (2009) most notably on the basis of the patchy phylogenetic distribution of *OS-ORF* and an observed association between the existence of intact *OS-ORF* sequences and genomic GC content. Finally, we examine three cases of predicted overlaps within the *INK4a* (Quelle et al. 1995), *XBP1* (Yoshida et al. 2001), and *GNAS1* (Klemke et al. 2001) genes in the human genome. In *XBP1*, the overlap is between two alternatively spliced exons whereas the overlaps in *INK4a* and *GNAS1* are thought to produce distinct proteins. Although these genes were reported to have complex regulatory properties (Chung et al. 2007), questions regarding their functionality remain (Nekrutenko and He 2006; Szklarczyk et al. 2007).

Methods

We simulated the evolution of overlapping genes as in Sabath et al. (2008b). In each run of the simulation, one gene was designated as known and the second as hypothetical. We examined the effects of the following factors on the ability of the two methods to detect selection in the hypothetical gene: (1) nonsynonymous/synonymous rate ratios in the hypothetical gene and the known gene (ω_h and ω_k, respectively), (2) overlap types (same-strand (SS) phases 1 and 2 and opposite-strand (OS) phases 0, 1, and 2), (3) sequence divergence (t), (4) sequence length, and (5) GC content. For each combination of overlap type, sequence divergence, and sequence length, we set ω_k and vary ω_h between 0.2 (strong purifying selection) and 1 (no selection). For each set of parameters, we generated 100 random pairs of overlapping genes. In order to test the influence of GC content, we generated random sequences with 30, 50, and 70% GC content and simulated their evolution with standard parameters (same-strand phase 1, length of 300 codons, $t = 0.4$, ω_k set to 0.2, and ω_h varied between 0.2 and 1). Sensitivity is defined as the percent of hypothetical genes under selection that were identified correctly by the method. Specificity is defined as the fraction of hypothetical genes that were incorrectly identified to be under selection when ω_h was set to 1 (i.e., no selection).

HSP70–OS-ORF Overlap

In addition to the 29 bacterial *HSP70* genes with an intact *OS-ORF* in Williams et al. (2009), we acquired 16 new sequences through BLASTing (Altschul et al. 1990) the nr database (limited to bacterial sequences) with the original sequences as queries. We constructed a multiple alignment of the sequences at the amino-acid level with CLUSTAW (Thompson et al. 2002) as implemented in the MEGA package (Kumar et al. 2008). Sites with gaps were removed resulting in a multiple-sequence alignment 561 codons in length. The multiple alignment is provided in the Supplementary Information. We then tested for selection on the *OS-ORFs* in all homologous pairs with *OS-ORF* sequence divergence of 50% or less.

Overlapping ORFs in the Human Genome

We identified the sequence of mouse *INK4a* (Accession: NM_001040654), based on Quelle et al. (1995) and obtained the orthologous gene sequences of 12 other species (human, rat, chimpanzee, orangutan, marmoset, macaque, cat, dog, cow, opossum, horse, and rabbit) through the UCSC Genome Browser (Karolchik et al. 2004). For the *GNAS1* overlap, we used eight sequences from human, chimpanzee, orangutan, gorilla, macaque, gibbon, colobus, and squirrel monkey (Nekrutenko et al. 2005). In addition, we obtained the orthologous mouse *GNAS1* CDS (accession AY519504) through a Blast search (Altschul et al. 1990). For the *XBP1* overlap, we obtained eight sequences from human, mouse, rat, cow, frog, chicken, zebrafish, and tilapia (Nekrutenko and He 2006) as well as three additional sequence from orangutan, gorilla, and pufferfish) through a Blast search (Altschul et al. 1990).

We constructed multiple alignments of each gene at the amino-acid level (see, Supplementary Information) with CLUSTAW (Thompson et al. 2002) as implemented in the MEGA package (Kumar et al. 2008), and removed sites with gaps. The length of the multiple alignment after removal of gaps was noted. GC content was measured as the mean across all sequences in each alignment. For each overlap, we calculated the sequence identity between human and mouse genes, at both the nucleotide and amino-acid levels, as a reference for comparisons among the three cases. Finally, we tested for selection on all pairs with sequence divergence of 50% or less. Runs in which the program did not converge (usually in pairs of highly similar sequences) were excluded from further analysis.

Results and Discussion

Simulation

We initially set the sequence length to 300 codons and $t = 0.4$, which corresponds to a sequence divergence of $\sim$12%. We set ω_k to 0.2 and varied ω_h between 0.2 and 1. The results are shown in Fig. 2a. Each square presents the results for SLG (solid blue: $P < 0.01$; dashed blue: $P < 0.05$) and FB (red) methods against ω_h (X-axis). An ideal detector is exemplified by a dashed green line. Each data point is the percentage of runs in which the methods

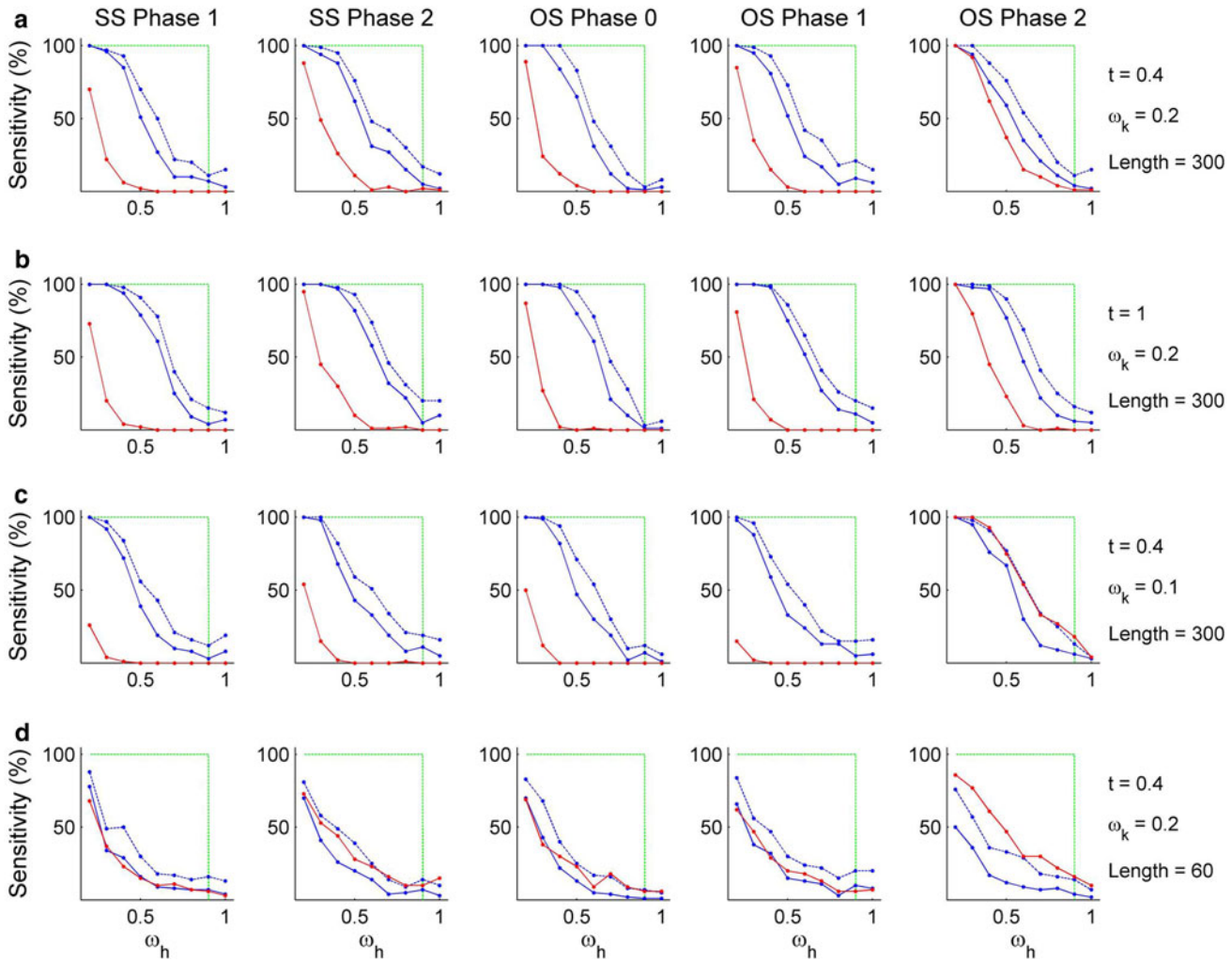

Fig. 2 Detection of selection by the SLG (*solid blue*: $P < 0.01$; *dashed blue*: $P < 0.05$) and FB (*red*) methods on simulated genes. An ideal detector is illustrated by a *dashed green* line. Each data point is the percentage of runs for which the methods detected selection. The five overlap types are shown in each column. Four sets of values for sequence divergence (t), ω_k, and sequence length are shown in each row (see text)

detected selection. The five overlap types are shown in each column. As expected, the sensitivity of both methods decreased with increase in ω_h. In all overlap types, SLG exhibits a higher sensitivity than FB, up to $\sim$80% in same-strand for $\omega_h = 0.4$. As expected, using SLG with P value of 0.05 (rather than 0.01), increase the method's sensitivity at the cost of lower specificity. For opposite-strand phase 2, both methods perform similarly. This phase is unique in that the third codon position of both genes corresponds, and thus, most changes are either nonsynonymous in both genes, or synonymous in both (Sabath et al. 2008b). This overlap phase was also reported to generate high rate of false-positive results (Firth and Brown 2006).

In the next three sets, we tested different values of t, ω_k, and sequence lengths, one parameter at a time. In Fig. 2b, we present the performance of the methods at high sequence divergence levels ($t = 1$, corresponding to a sequence divergence of $\sim$24%). For both methods, the results are similar to those at low sequence divergence. In Fig. 2c, we present the results for stronger selection level on the known gene ($\omega_k = 0.1$). The performance of SLG is similar to that in (a) and (b), whereas the sensitivity of FB is reduced in same-strand phases 1 and 2 and opposite-strand phases 0 and 1. In Fig. 2d, we present the results for short sequence length (60 codons). Under these conditions, SLG and FB perform similarly, with SLG showing reduced sensitivity compared to (a), (b), and (c). Finally, we tested whether the methods are influenced by GC content. We tested sequences with 30, 50, and 70% GC content and simulated their evolution with standard parameters (see Methods). We find that both methods are highly robust to differences in GC content (Supplementary Fig. S1).

Overall, the simulation demonstrates that, under most conditions, SLG performance is as good as FB or higher.

The advantage of using SLG over FB is that it performs better when the known gene is under strong purifying selection, whereas both methods perform alike on short sequences. In addition, SLG was found to be more robust to overlap type in comparison to FB, whose performance is more variable, especially in the case of opposite-strand phase 2 overlaps. Similarly to FB, SLG can be applied to multiple sequences by choosing only neighboring terminal taxa in the phylogenetic tree (Firth and Brown 2006). This approach, while ingenious, only indirectly addresses the phylogenetic framework and may be biased for trees with non-uniform branch-length distribution. In future studies, it would be beneficial to take full advantage of the maximum-likelihood framework that allows testing hypotheses concerning variable selection pressures among lineages and sites (Nielsen and Yang 1998; Zhang et al. 2005). This might be of special significance for overlapping genes because they may exist as a non-functional ORFs before they become functional (Keese and Gibbs 1992).

Testing the Functionality of Functionality of *OS-ORF*

The functionality of *HSP70* and its overlapping *OS-ORF* constitute a central tenet of the hypothesis concerning the origin of the two aaRS classes (Carter and Duax 2002; Rodin and Ohno 1995). A recent study by Williams et al. (2009) cast doubt on the functionality of *OS-ORF*. We used SLG to determine whether or not selection operates on *OS-ORF*. We identified 45 bacterial *HSP70* genes with an intact *OS-ORF* and tested for selection on the *OS-ORFs* in all homologous pairs with *OS-ORF* divergence of up to 50%. The results are shown in Fig. 3. For each pair, the amino-acid sequence divergence of the *OS-ORFs* was plotted against that of *HSP70*. Pairs for which the method did not detect selection (at *P* values larger then 0.01) are marked in blue, and pairs for which a signature of selection was found are marked in red. First, in all pairs the divergence between the *OS-ORF* sequences is higher than the divergence of *HSP70* (up to 30% higher), suggesting that *HSP70* is under much stronger selection pressure. Second, pairs for which a signature of selection was found are concentrated in the upper divergence range of *OS-ORF*. The detection of selection signatures only in highly divergent pairs suggests that these are false-positive results and that *OS-ORF* is not a functional gene.

The most likely reason for inaccuracy at high sequence divergence values is that the method estimates the probability of one codon to change to another by summing over all possible paths. With the increase in divergence, the number of possible paths raises and, consequently, the power of the method to recover to true path decreases. In addition, the reduced quality in alignment of distant sequences could also contribute to biased inference of

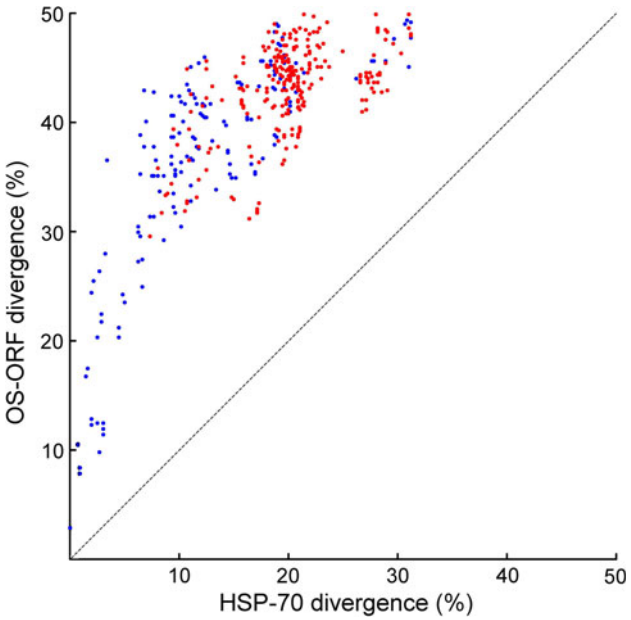

Fig. 3 Testing for selection on *OS-ORF*. The amino-acid sequence divergence of the *OS-ORF* is plotted against that of *HSP70* for all pairs of homologous sequences up to 50% divergence of the *OS-ORF*. (*Red*) Sequence pairs in which the method detected selection ($P < 0.01$). (*Blue*) Sequence pairs in which the method did not detect selection

selection (Schneider et al. 2009). These results imply that ~30% divergence between sequences should be the upper boundary for using the SLG method. This boundary is comparable to the boundary of 25% divergence suggested for exon detection using a single-coding genes analogous method (Nekrutenko et al. 2002).

Overlapping ORFs in the Human Genome

We chose *INK4a*, *XBP1*, and *GNAS1* because they are considered to be the best examples of overlaps in the human genome (Chung et al. 2007). Since the aim of this study is to evaluate our method, we thought to concentrate on sequence-based evidence and not to review the full body of experimental evidence for and against the functionality of each gene. We, therefore, do not make any a priori assumption about the functionality of any of the genes and test for selection on both overlapping ORFs. The results (Fig. 4) are shown in the same format as in Fig. 3 where the amino-acid sequence divergence of an ORF was plotted against that of its overlapping ORF. Pairs for which the method did not detect selection (At *P* values larger then 0.01) are marked in blue, and pairs for which a signature of selection was found are marked in red. In each plot, the ORF under examination is highlighted in green. For example, in Fig. 4a, the top plot presents the results of testing *ARF* and the bottom plot presents the results of

testing *INK4a*. In order to complement our method and to allow comparison among the three overlaps, we also note three benchmark genomic properties, GC content and sequence identity at both the nucleotide and amino-acid levels. The average GC content of human genes is 52%, and the median identity between human and mouse coding sequences is 85% at the DNA level and 78.5% at the amino-acid level (Waterston et al. 2002). High GC content, which could be the result of several mutational processes (Graur and Li 2000), leads to reduction in the frequencies of stop codons (which are AT rich) in overlapping ORFs and, hence, contributes to the conservation of overlapping ORFs even in the absence of selection (Sabath et al. 2008a; Silke 1997; Williams et al. 2009). In Table 1, we note: (1) overlap length after excluding gap-containing sites, (2) GC content, (3) percent identity between human and mouse genes at the DNA level, (4) percent identity between human and mouse genes at the amino-acid level, (5) number of pairs under 30% divergence, in which the method detected selection ($P < 0.01$), (6) total number of pairs under 30% divergence, and (7) percentage of positives out of the total number of pairs.

INK4a/ARF Overlap

The *INK4a/ARF* overlapping genes are known as regulators of tumor suppression pathways (reviewed in Szklarczyk et al. 2007). Using our method, we find considerable evidence of selection (80.3% of the pairs, Fig. 4a, bottom plot) acting on *INK4a*, whereas *ARF* exhibits very weak signature of purifying selection (2.5%, Fig. 4a, top plot). All sequence pairs exhibit equal or higher amino-acid divergence at the *ARF* frame than at the *INK4a* frame as seen by all of the points being located either on the diagonal or above it (Fig. 4a), similarly to the *HSP70/OS-ORF* overlap (Fig. 3). The overlap region has a very high GC content of 74% (Table 1). The human-mouse amino-acid sequence identity of *INK4a* is average (76.6%), while that of *ARF* is considerably low (50.8%). It is possible that our method failed to detect selection on ARF because of the

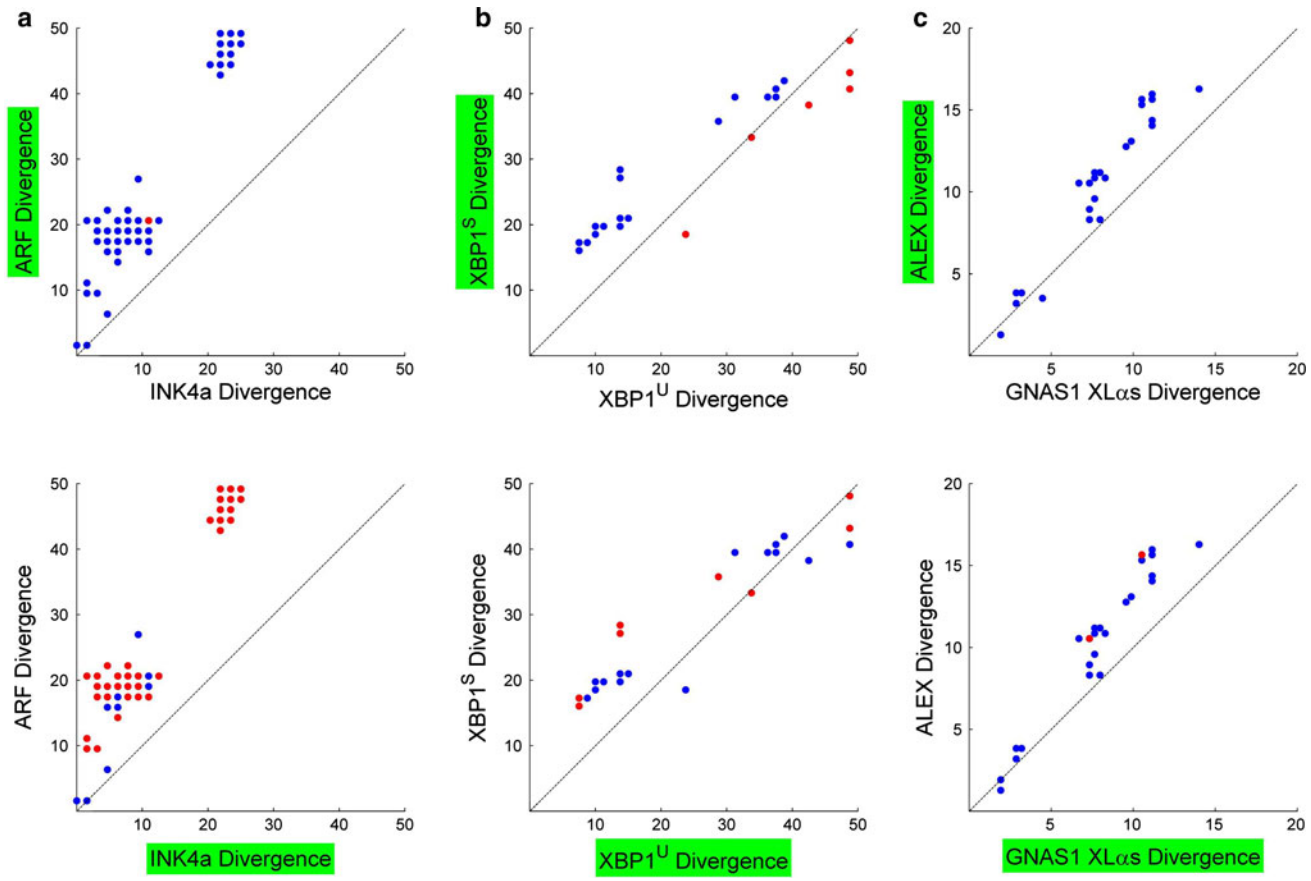

Fig. 4 Testing for selection on *INK4a/ARF* (**a**), *XBP1* (**b**), and *GNAS1/ALEX* (**c**). The amino-acid sequence divergence of each ORF is plotted against that of its overlapping ORF for all pairs. In each plot, the ORF, which is tested for selection is highlighted in *green*. For example, in **a**, the *top plot* presents the results of testing *ARF* and the *bottom plot* presents the results of testing *INK4a*. (*Red dots*) Sequence pairs in which the method detected selection ($P < 0.01$). (*Blue dots*) Sequence pairs in which the method did not detect selection

Table 1 Testing for selection on overlaps within three human genes

Locus	Length (codons)	GC (%)	H-M NT (%)[a]	H-M AA (%)[b]	ORF	Positive[c]	Total[d]	%[e]
INK4a/ARF	64	74	77.6	50.8	ARF	1	40	2.5
				76.6	INK4a	57	71	80.3
XBP1	85	53	92.2	91.7	$XBP1^U$	5	13	38.5
				82.1	$XBP1^S$	1	12	8.3
GNAS1/ALEX	314	69	68.3	47.2	ALEX	0	26	0
				46.4	GNAS1 XLαs	2	26	7.7

[a] Percent identity between human and mouse genes at the DNA level

[b] Percent identity between human and mouse genes at the amino-acid level

[c] Number of pairs under 30% divergence, in which the method detected selection ($P < 0.01$)

[d] Total number of pairs under 30% divergence

[e] Percentage of positives out of the total number of pairs

short length of the region. However, together with the amino-acid divergence and high GC content, these results imply that the overlapping region of *INK4a* is functional whereas that of *ARF* is most likely spurious. This conclusion is in agreement with the absence of this overlap in chicken (Kim et al. 2003) and the lack of apparent function of *ARF* in other species. Arguing for functionality of *ARF*, Szklarczyk et al. (2007) have simulated random sequences corresponding to the amino-acid sequence of *INK4a* using the codon usage of all human genes. They found that only 4% of the simulated sequences contained full ORF in the *ARF* frame, suggesting that selection must be preserving the ORF. However, local deviation of codon usage (as reflected by the high GC content in the *INK4a/ARF* overlap) could be the result of both selection and mutation. Therefore, it is possible that the absence of stop codons disturbing the *ARF* sequence is simply the result of mutational bias rather than selection acting to conserve its sequence.

$XBP1^U/XBP1^S$ overlap

XBP1 protein is associated with regulation of the expression of genes implicated in the unfolded protein response (Nekrutenko and He 2006). We find weak but significant evidence for selection (38.5% of the pairs) acting on $XBP1^U$, and a weaker signal (8.3%) on $XBP1^S$. Unlike *HSP70/OS-ORF* and *INK4a/ARF* overlaps, the divergence of sequence pairs do not have a distinct pattern and the points are scattered on both sides of the diagonal (Fig. 4b). The overlap region has a moderate GC content of 53% (Table 1), suggesting low mutational bias. The human–mouse amino-acid sequence identity of both ORFs is higher than average (91.7 and 82.1% for $XBP1^U$ and $XBP1^S$, respectively), suggesting that the region is under strong purifying selection. Based on these results, we predict that $XBP1^U$ is functional with a high degree of confidence. The functionality of $XBP1^S$ is not so well supported by our method. However, given the short length of the overlap (which makes the detection of selection difficult), the phylogenetic-wide conservation of this ORF (Nekrutenko and He 2006), and the moderate GC content of this region, functionality cannot be ruled out.

GNAS1/ALEX Overlap

The *GNAS1* locus encodes the alpha subunit of the stimulatory G protein of adenylyl cyclase (Kozasa et al. 1988; Levine et al. 1991). This protein is a member of the G proteins, a family of signal-coupling proteins that mediate numerous transmembrane hormonal and sensory transduction processes (Bourne et al. 1990). This gene is also known to have complex imprinting patterns (Hayward et al. 1998). We focused on the *GNAS1 XLαs* exon and its overlapping ORF, *ALEX*. We find very little evidence of selection (7.7% of the pairs) acting on *GNAS1 XLαs* exon, and no signal (0%) of selection on *ALEX*. The overlap region has a high GC content of 69% (Table 1). The human–mouse amino-acid sequence identity of both ORFs is lower than average (46.4 and 47.2% for *GNAS1 XLαs* and *ALEX*, respectively). Surprisingly, the sequence identity at the nucleotide level (68.3%) is much higher than the amino-acid sequence identity of both ORFs. This difference between the identities at the nucleotide and amino-acid levels (>20%) is unique to the *GNAS1* overlap, whereas in the other two overlaps the minimum difference is $\sim$1% (Table 1). The low amino-acid conservation of the *GNAS1 XLαs* exon is in striking contrast to the rest of the protein, which is nearly identical between the two species (Supplementary Fig. S2). Unlike *INK4a/ARF* and *XBP1* overlaps, the *XLαs/ALEX* overlap is 314 long, a length at which the sensitivity of our method is high (Fig. 2). Based on these results, we do not find strong support for the functionality of neither *XLαs* nor *ALEX*. The only observation that remains in support of both overlapping ORFs

being functional is that none of the multiple insertions and deletions in this region disturb the reading frame (Supplementary Fig. S2; Nekrutenko et al. 2005). In order to evaluate the significance of this observation, it would be necessary to apply an evolutionary model of insertions and deletions.

Nekrutenko et al. (2005) interpreted the high, but relatively similar, evolutionary rate of both ORFs as "constant adjustment between the two proteins aimed at maintaining mutual affinity." Furthermore, Nekrutenko et al. (2005) suggested a model, in which purifying selection is acting on the two proteins to maintain both ORFs and their expression coupling. Using our method, we show that there is no strong purifying selection acting on these proteins (only a weak selection on *XLαs* exon if any). Once we exclude the scenario of purifying selection acting on both proteins, their equal rate of evolution could simply be explained by drift, which does not distinguish between two non-functional ORFs.

We studied the performance of our new method for the detection of functional overlapping genes. By simulation, we compared the method to the FB method and tested both methods across different overlap types. We found that under most conditions, our method predicts functionality with higher sensitivity while maintaining high specificity. Finally, we examined the overlaps within *HSP-70*, *INK4a*, *XBP1*, and *GNAS1* genes. We conclude that long overlaps in non-viral genomes should be treated with caution since many of them may be spurious.

Acknowledgments We thank Dr. Anton Nekrutenko and an anonymous reviewer for their useful comments. This work was supported in part by US National Library of Medicine Grant LM010009-01 to Dan Graur and Giddy Landan.

References

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

Bourne HR, Sanders DA, McCormick F (1990) The GTPase superfamily: a conserved switch for diverse cell functions. Nature 348:125–132

Carter CW, Duax WL (2002) Did tRNA synthetase classes arise on opposite strands of the same gene? Mol Cell 10:705–708

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

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

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. Nucl Acids Res 27:4636–4641

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

Graur D, Li W-H (2000) Fundamentals of molecular evolution. Sinauer Associates, Sunderland, MA

Hayward BE, Kamiya M, Strain L, Moran V, Campbell R, Hayashizaki Y, Bonthron DT (1998) The human GNAS1 gene is imprinted and encodes distinct paternally and biallelically expressed G proteins. Proc Natl Acad Sci USA 95:10038–10043

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ (2004) The UCSC Table Browser data retrieval tool. Nucl Acids Res 32:D493–D496

Keese PK, Gibbs A (1992) Origins of genes: "big bang" or continuous creation? Proc Natl Acad Sci USA 89:9489–9493

Kim SH, Mitchell M, Fujii H, Llanos S, Peters G (2003) Absence of p16INK4a and truncation of ARF tumor suppressors in chickens. Proc Natl Acad Sci USA 100:211–216

Klemke M, Kehlenbach RH, Huttner WB (2001) Two overlapping reading frames in a single exon encode interacting proteins—a novel way of gene usage. EMBO J 20:3849–3860

Konstantopoulou I, Ouzounis CA, Drosopoulou E, Yiangou M, Sideras P, Sander C, Scouras ZG (1995) A Drosophila hsp70 gene contains long, antiparallel, coupled open reading frames (LAC ORFs) conserved in homologous loci. J Mol Evol 41:414–420

Kozasa T, Itoh H, Tsukamoto T, Kaziro Y (1988) Isolation and characterization of the human Gs alpha gene. Proc Natl Acad Sci USA 85:2081–2085

Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform 9:299–306

Lavorgna G, Dahary D, Lehner B, Sorek R, Sanderson CM, Casari G (2004) In search of antisense. Trends Biochem Sci 29:88–94

Levine MA, Modi WS, O'Brien SJ (1991) Mapping of the gene encoding the alpha subunit of the stimulatory G protein of adenylyl cyclase (GNAS1) to 20q13.2–q13.3 in human by in situ hybridization. Genomics 11:478–479

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, 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

Monnerjahn C, Techel D, Mohamed SA, Rensing L (2000) A non-stop antisense reading frame in the grp78 gene of Neurospora

crassa is homologous to the Achlya klebsiana NAD-gdh gene but is not being transcribed. FEMS Microbiol Lett 183:307–312

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, Makova KD, Li WH (2002) The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. Genome Res 12:198–202

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

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936

Palleja A, Harrington ED, Bork P (2008) Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? BMC Genomics 9:335

Quelle DE, Zindy F, Ashmun RA, Sherr CJ (1995) Alternative reading frames of the INK4a tumor suppressor gene encode two unrelated proteins capable of inducing cell cycle arrest. Cell 83:993–1000

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

Rodin SN, Ohno S (1995) Two types of aminoacyl-tRNA synthetases could be originally encoded by complementary strands of the same nucleic acid. Orig Life Evol Biosph 25:565–589

Rother KI, Clay OK, Bourquin JP, Silke J, Schaffner W (1997) Long non-stop reading frames on the antisense strand of heat shock protein 70 genes and prion protein (PrP) genes are conserved between species. Biol Chem 378:1521–1530

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

Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D (2009) Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. Genome Biol Evol 1:114–118

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

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

Thompson JD, Gibson TJ, Higgins DG (2002) Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinform Chapter 2: Unit 2.3

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

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

Yoshida H, Matsui T, Yamamoto A, Okada T, Mori K (2001) XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. Cell 107:881–891

Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22:2472–2479