Rapid asymmetric evolution of a dual-coding tumor suppressor INK4a/ARF locus contradicts its function

Radek Szklarczyk*†, Jaap Heringa*, Sergei Kosakovsky Pond‡, and Anton Nekrutenko§¶

*Centre for Integrative Bioinformatics, Vrije University, De Boelelaan 1081a, 1081HV, Amsterdam, The Netherlands; ‡Department of Pathology, University of California at San Diego, La Jolla, CA 92093; and §Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, University Park, PA 16803

Edited by Russell F. Doolittle, University of California at San Diego, La Jolla, CA, and approved June 14, 2007 (received for review April 7, 2007)

INK4a/ARF **tumor suppressor locus encodes two protein products, INK4a and ARF, essential for controlling tumorigenesis and mutated in more than half of human cancers. There is no resemblance between the two proteins: their coding regions are assembled by alternative splicing of two mutually exclusive 5′ exons into a constitutive one containing overlapping out-of-phase reading frames. We show that the dual-coding arrangement conflicts with the high cost of mutations within *INK4a/ARF*. Unexpectedly, the locus evolves rapidly and asymmetrically, with ARF accumulating the majority of amino acid replacements. Rapid evolution drives both INK4a and ARF proteins out of sync with other members of the RB and p53 tumor suppressor pathways, both of which are controlled by the locus. Yet, the asymmetric behavior may be an intrinsic property of dual-coding exons: INK4a/ARF closely mimics the evolution of 90 newly identified genes with similar dual-coding structure. Thus, the strong link between mutations in INK4a/ARF and cancer may be a direct consequence of the architecture of the locus.**

CDKN2A | p53 | retinoblastoma protein | overlapping protein-coding regions | cancer evolution

INK4a/ARF is a unique locus encoding two structurally unrelated proteins via overlapping coding regions. Its products regulate independent tumor suppression pathways and strike the fine balance between cancer and physiological aging (1). At first glance, the locus is unassuming: it is compact (at a little more than 25 kbp, it is shorter than most human genes), contains four exons, and produces just two ubiquitously expressed transcripts (Fig. 1). The transcripts originate at two different upstream exons, 1α and 1β, but share an internal dual-coding exon 2 (see Fig. 1*A*). The transcript originating at exon 1α is commonly called INK4a (inhibitor of cyclin D-dependent kinases CDK4 and CDK6) and is translated to a protein called $p16^{INK4a}$. Its counterpart beginning with exon 1β is referred to as ARF (alternative reading frame) and is translated into polypeptide $p14^{ARF}$ in human and $p19^{ARF}$ in mouse. Because they are translated from alternative reading frames of exon 2 (Fig. 1 and ref. 2), the two proteins share no sequence similarity. *INK4a/ARF* is one of only three human genes that have been experimentally confirmed to produce proteins from overlapping reading frames. The locus is also unique in that the choice of the translated reading frame is determined by alternative splicing, whereas in the other two genes, *GNAS1* and *XBP1*, translation is rerouted between the two overlapping reading frames either by a stochastic choice of alternative start codons (3, 4), or by endonucleolytic cleavage (5) (Fig. 2).

Further complicating matters is the fact that the two protein products initiate independent tumor suppression programs. INK4a maintains RB, a tumor suppressor that prevents cells from entering the S phase and initiating DNA replication (thus limiting cell proliferation), in its active nonphosphorylated state by inhibiting cyclin D-dependent kinases CDK4 and CDK6. ARF, on the other hand, safeguards tumor suppressor activity of the p53 pathway by binding and inhibiting one of its chief antagonists: MDM2, a ubiquitine ligase (1, 6–8). Besides MDM2, ARF interfaces with a number of other proteins, of which the interaction with nucleophosmin (NPM or B23) is best studied (9, 10). The level of INK4a expression increases with age. Because INK4a is a potent inhibitor of cell-cycle progression, its elevated expression diminishes the regenerative potential of tissues and contributes to aging. This increase in expression, in turn, reduces the probability of cancer-causing hyperproliferation, thus modulating the balance between regeneration and cancer (11–13).

The complexity and physiological importance of functions controlled by *INK4a/ARF* seem to be at odds with the structural organization of the locus. Overlapping reading frames (dual-coding) create a strong codependency between the proteins: a synonymous change in one frame often leads to an amino acid replacement in the other. The high intrinsic mutation rate of the locus amplifies this effect. *INK4a/ARF* is one of the most frequently mutated loci in human cancers, with the frequencies ranging from 30% in esophageal tumors to 100% in pancreatic carcinomas (14). Overall, in various cancers, 206 nucleotides in the overlap between the two reading frames (Fig. 1) carry combinations of 117 point mutations, of which 40, 15, and 62 affect INK4a, ARF, or both proteins, respectively (15, 16). What is the evolutionary benefit of maintaining overlapping reading frames in *INK4a/ARF*, especially in light of the paucity of dual-coding genes in eukaryotic genomes? To answer this question, we first studied molecular evolutionary dynamics of the locus in sequenced mammalian genomes. Second, we compared the evolutionary rates in the INK4a and ARF reading frames to those of other members of the RB and p53 tumor suppression pathways. Finally, we analyzed a comprehensive collection of human and mouse transcripts in pursuit of genes with a structure similar to that of *INK4a/ARF*.

Results and Discussion

Dual-Coding Region Is Conserved in Mammalian Ink4A/ARF. To study the evolution of the *INK4a/ARF* locus, we constructed an alignment of orthologous protein-coding sequences from seven mammalian species: human, chimpanzee, macaque, mouse, rat, dog, and cow. To contrast the selective regimes between single- and dual-coding regions of the locus, we partitioned the alignment into segments corresponding to three distinct regions of the locus: 1α, 1β, and 2

Author contributions: A.N. designed research; R.S. and A.N. performed research; J.H. and S.K.P. contributed new reagents/analytic tools; R.S., S.K.P., and A.N. analyzed data; and R.S. and A.N. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Abbreviations: dN, nonsynonymous; dS, synonymous; dN^{alt}, nonsynonymous alternative reading frame; DC, dual coding; SC, single coding.

†Present address: Center for Molecular and Biomolecular Informatics, Nijmegen Center for Molecular Life Sciences, Radboud University Nijmegen Medical Center, Geert Grooteplein 28, 6525 GA, Nijmegen, The Netherlands.

¶To whom correspondence should be addressed at: 505 Wartik Laboratory, Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, University Park, PA 16802. E-mail: anton@bx.psu.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0703238104/DC1.

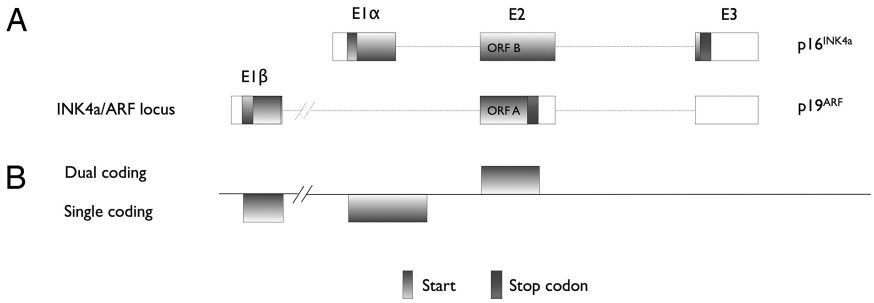

Fig. 1. *INK4a/ARF* locus resides on chromosome 9p21 and spans ≈25 kb. (*A*) Exon/intron structure of two ubiquitously expressed and one tissue-specific protein products. Exon sizes are 150 bp for exon 1α, 316 bp for exon 1β, and 307 bp for exon 2 (206 bp are translated in two reading frames). (*B*) Locations of single- and dual-coding regions of the gene.

(Fig. 1*B*). This notation will be used throughout the rest of the paper.

Biochemical studies of INK4a and ARF point to an unequal contribution of amino acids encoded within the overlap (exon 2) to the functions of the two proteins. Whereas the functionality of the exon 2-encoded INK4a portion is well established (8), it is less clear for the ARF polypeptide. A prevailing view is that only the exon 1β-encoded portion of ARF is biologically relevant and the exon 2-encoded remainder has no known function. This view is supported by experimental data from mice (17) and the fact that chicken lacks the region of overlap altogether, with ARF encoded only by exon 1β (18). The only apparent contradiction to this view is furnished by the data from the human ARF, where a nucleolar localization domain, a key functional element of ARF, is encoded within the overlap region of exon 2 (19). However, the disruption of exon 2 manifests itself in a reduction of ARF activity that can be overcome by an increase in the expression level (6). Despite these conflicting observations, sequence analyses tip the scales toward the likely functionality of the overlap-encoded ARF segment. If the C terminus of ARF were nonfunctional, it could be terminated by a stop codon at any point within exon 2. However, in the seven mammalian species considered here, the length of the overlap between *INK4a* and *ARF* reading frames remains very nearly constant (67–68 codons). Alternatively, the overlap can be an artifact of nucleotide composition dictated by codon usage in the *INK4a* reading frame of exon 2. To evaluate this possibility, we performed a permutation test (20). We simulated 1 million exon 2 sequences by randomly drawing codons corresponding to each amino acid encoded by the *INK4a* frame from the codon usage table compiled from all known human genes. As a result, all simulated sequences encoded the same polypeptide in the *INK4a* frame but had different nucleotide sequences because of the random sampling of synonymous codons. Finally, we translated simulated sequences in the *ARF* frame and counted all those that did not contain premature stop codons. Only ≈4% of simulated sequences contained full length ARFs [supporting information (SI) Fig. 6], suggesting that the section of the *ARF* frame within the overlap is likely functionally relevant.

Rapid Asymmetric Evolution of the Locus. Products of the *INK4a/ARF* gene control two key tumor suppression pathways. Intuitively, a locus providing such an essential functionality should be under strong selective constraint. However, the opposite is true: both reading frames evolve significantly faster than the majority of other protein-coding regions, with the ARF frame leading that race. Pairwise nucleotide divergence (measured in expected substitutions per 100 sites) between the three regions of the human *INK4a/ARF* locus and its mammalian counterparts is unusually high (SI Table 2): the most conserved region of the *INK4a/ARF* locus, exon 2, is ≈26% (95% profile likelihood confidence interval: 19–36%) divergent between human and mouse at the nucleotide level: a sharp increase from the 18% median estimated for ≈8,500 pairs of human/mouse orthologous protein-coding regions (see *Materials and Methods*). Divergence levels of exons 1α and 1β between human and mouse are still higher: 33% (23–47%) and 46% (37–57%), respectively. To understand the effect of high rates of nucleotide substitution on the evolution of encoded proteins, we initially estimated rates of silent [synonymous (dS)] and amino acid changing [nonsynonymous (dN)] substitution for the *ARF* and *INK4a* frames within the three partitions (Table 1). The *INK4a* frame seems to accumulate the bulk of synonymous substitutions. Both exons exhibit an abundance of synonymous substitutions, suggestive of strong purifying selection on this gene and high mutation rates in the gene, as corroborated by a large proportion of human–mouse orthologs with significantly lower dS than that in either exon ($P = 0.05$, likelihood ratio tests with multiple testing correction; see *Materials and Methods*). These proportions were 53% (4,490/8,467) for exon 1α and 75% (6,300/8,467) for exon 2/INK4a (Fig. 3). The evolutionary history of *ARF*, on the other hand, seems to be quite the opposite. A comparison with other genes reveals an unusually high dN: 100% (8,467/8,467) of genes have significantly lower dN for exon 1β and 99.98% (8,465/8,467)

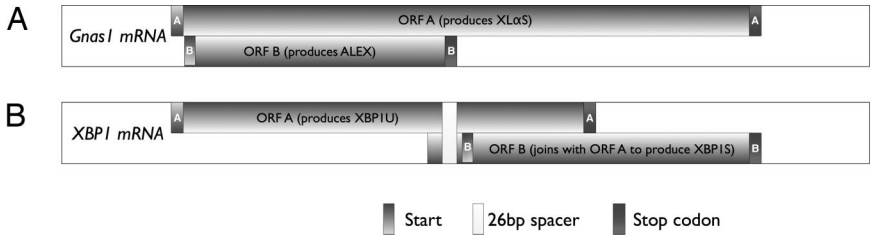

Fig. 2. In addition to *INK4a/ARF*, two other mammalian genes produce empirically confirmed proteins from overlapping reading frames: GNAS1 and XBP1. (*A*) A transcript of GNAS1 gene contains two reading frames and produces two structurally unrelated proteins XLαs and ALEX by differential utilization of translation start sites. (*B*) A newly transcribed *XBP1* mRNA can only produce protein XBP1U from ORF A. Removal or a 26-bp spacer (yellow rectangle) joins the beginning of ORFA with ORFB and translates into a different product called XBP1S.

Table 1. Substitution rates at the *INK4a* locus in a human–mouse comparison

Frame/Exon	Codons	dS	dN	ω = dN/dS	Neutrality tests
INK4a/1α	50	1.14 (0.5–∞)	0.18 (0.11–0.32)	0.14 (0.003–0.37)	$\omega < 1$: $P < 0.01$
INK4a/2	68	1.15 (0.63–∞)	0.13 (0.07–0.23)	0.09 (0.02–0.20)	$\omega < 1$: $P < 0.01$
ARF/1β	105	0.51 (0.34–0.81)	0.46 (0.37–0.63)	0.86 (0.50–1.59)	$\omega \neq 1$: $P = 0.73$
ARF/2	68	0.08 (0.0–0.15)	0.41 (0.28–0.67)	7.54 (3.2–∞)	$\omega > 1$: $P < 0.01$

All rates are estimated by maximum likelihood using the MG94xHKY85 codon substitution model. Indicated 95% confidence intervals and *P* values for the neutrality tests (for given alternatives) are estimated by using parametric bootstrap based on 100 replicates.

for *exon2/ARF*. We caution, however, that these rates cannot be readily interpreted in the region of overlap (exon 2), because of the rates of evolution in one frame depend on the context in the other (21, 22). Even so, this behavior is drastically different from the evolutionary regimes of two other mammalian genes with overlapping reading frames, *GNAS1* and *XBP1*, where both reading frames seem to rapidly coevolve with dN rates of overlapping reading frames being approximately equal (20, 21). Thus, it seems that substitution events within a single stretch of genomic DNA result in discordant selective pressures on the reading frames on the *INK4a/ARF* locus, a pattern previously reported only for short terminal overlaps in bacterial and viral genomes (23).

The *INK4a/ARF* locus is a member of INK4 cell cycle inhibitor gene family, which, in higher vertebrates (mammals and birds), includes four loci: *INK4a/ARF*, *INK4b*, *INK4c*, and *INK4d* (reviewed in ref. 8). Although *INK4a/ARF* and *INK4b* originated from a single duplication event (24), their evolutionary paths are different: the former accumulates nonsynonymous changes almost three times faster than the latter (dN = 0.12 vs. 0.04 for *INK4a/ARF* and *INK4b* exon 2, respectively).

INK4a and ARF Evolve Disproportionally to Their Pathway Counterparts. INK4a and ARF interact with a number of other proteins and serve as nodes in independent RB and p53 tumor suppression pathways (Fig. 4). To maintain functionality, pathway elements must coevolve to maintain mutual interaction, a well documented property that can be used to identify interacting partners (25–27). To check how the rapid evolution of INK4a and ARF fits within the frameworks of RB and p53 pathways, we collected orthologous protein-coding regions for each pathway element (shown in Fig. 4) and reconstructed gene trees using codon evolutionary models [for this analysis, we used a simplified version of the RB and p53 pathways as described in (6–8)]. To compare evolutionary rates among members of each pathway, we applied a series of relative ratio tests (28) (see *Materials and Methods* and Fig. 4) that yielded a number of unexpected conclusions. In the p53 pathway, the rate of ARF evolution is substantially higher than that of its interacting partners. Importantly, if we assume that only the exon 1β-encoded portion of the ARF polypeptide is functional, it still accumulates nonsynonymous changes five times faster than MDM2. The situation is even more dramatic for the RB pathway, where nonsynonymous substitution rates within INK4a-encoding exons 1α and 2 are either much higher compared with CDK4 and CDK6 or nonproportional (e.g., between exon 1α and CDK4), suggesting lineage-specific selection. However, we cannot exclude the possibility that only a small number of interacting sites coevolve among the interacting partners, which may not be reflected in the comparison of gene trees (29).

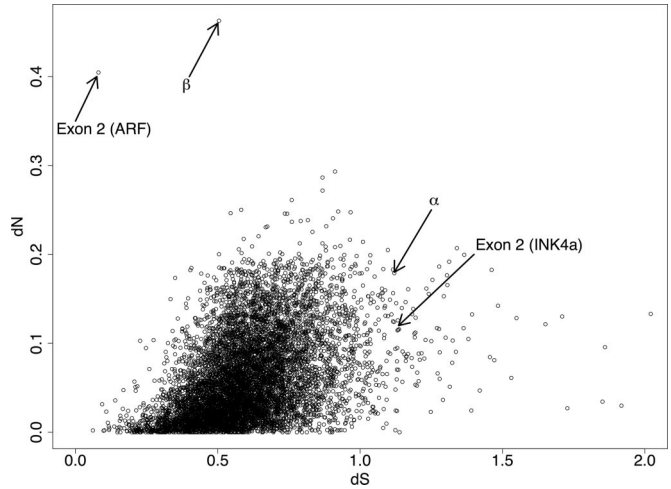

Fig. 3. Comparison of nonsynonymous and synonymous evolution rates of ≈8,500 pairs of human/mouse orthologous protein-coding regions (background divergence level) to the rates within INK4a/ARF exons.

More Dual Coding in Human Genes. Are there more *INK4a/ARF*-like loci in our genome? A recent attempt (30) to identify human genes with alternatively spliced dual-coding exons did not identify the *INK4a/ARF* locus. In addition, the study concluded that the dual-coding property of exons is rarely conserved across mammals. *INK4a/ARF* clearly contradicts this hypothesis. In an attempt to reconcile these conflicting observations, we used splicing graphs recently compiled for human transcriptome (31) to identify genes structurally similar to *INK4a/ARF*. We were looking specifically for genes that produced transcripts using two reading frames within a dual-coding exon (see SI Fig. 7 and *Materials and Methods*). Throughout the following discussion, we refer to the two reading frames of a dual-coding exon as constitutive and alternative. A constitutive frame is used in the major transcript of the gene (this frame is typically the frame annotated in sequence databases), whereas an alternative frame is used in minor, less abundant transcripts. In the case of *INK4a/ARF*, this assignment was arbitrary: *INK4a* became constitutive and *ARF* an alternative frame. After identifying a dual-coding exon in a human transcript, we checked (*i*) whether the splicing events that switch between constitutive and alternative reading frames were conserved in both human and mouse splicing graphs; and (*ii*) that the two reading frames were intact across mammalian species. Ninety dual-coding exons satisfied both conditions. Suprisingly, only 9 of 185 regions reported in ref. 30 qualified as dual-coding by our method. The low degree of overlap between the two data sets is largely due to the stringent requirement for cross species conservation of both splicing events and reading frame length conservation. Approximately 90% of dual-coding regions in our set span one or two exons. We further classified types of splicing events leading to a reading frame switch. The most common splicing event, exon skipping, constituted 57% of all events. This exon skipping was followed by approximately equal numbers of alternative donor/acceptor splice sites and intron retention events (≈14% each). Twice as many regions as expected under the uniform distribution begin at the 3′ terminal of the gene (last 5% of the gene), and half as many are located at the beginning

EVOLUTION

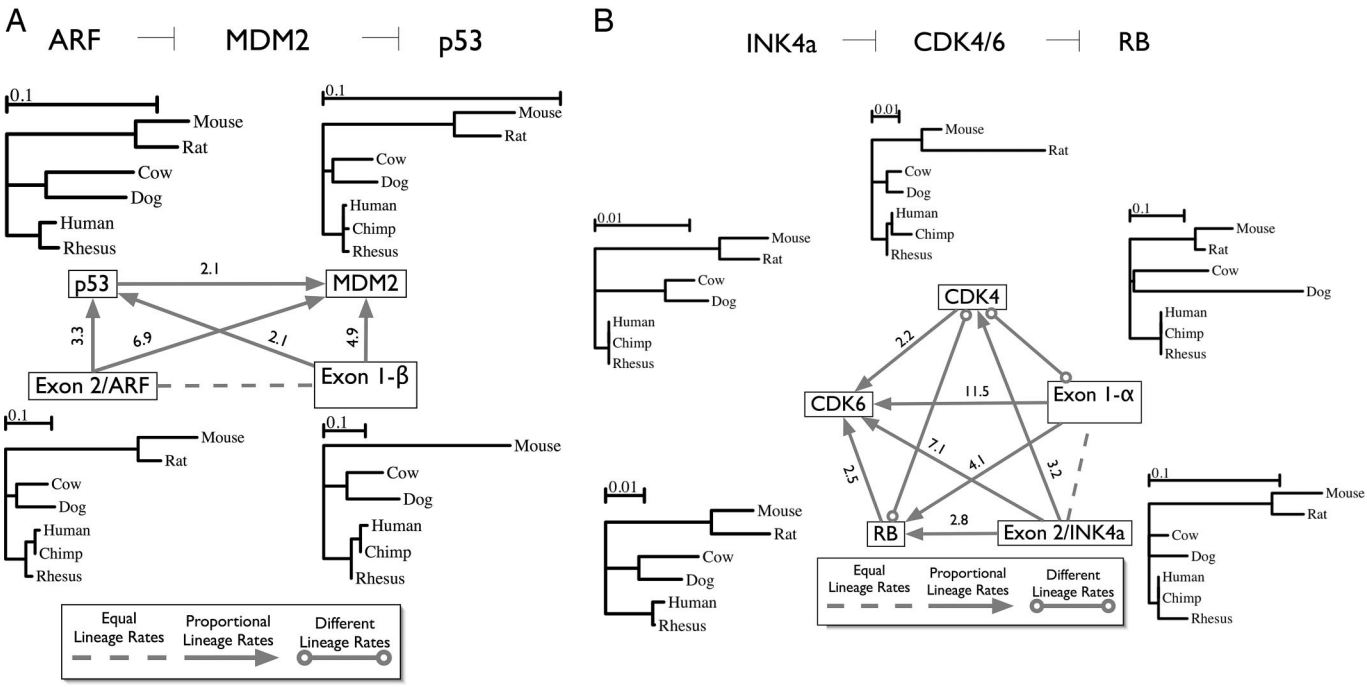

Fig. 4. Comparison of nonsynonymous substitution rates among members of RB and p53 signaling pathways. Unrooted trees are scaled on dN estimated by maximum likelihood under a codon substitution model. Connections between genes describe evolutionary relationships between sequences. Labels above denote the relative rate of accumulation of changes (e.g., exon 2 gains nonsynonymous mutations 3.2 times faster than its interacting gene, CDK4, whereas the rates of amino acid substitutions between the two exons of INK4a are equal). (*A*) The evolution of ARF-MDM2-p53 pathway. ARF safeguards tumor suppressor activity of p53 by binding and inhibiting one of its antagonists: MDM2. (*B*) INK4a-CDK4/6-RB pathway. INK4a maintains RB, a tumor suppressor that prevents cells from entering the S phase and initiating DNA replication, in active nonphosphorylated state by inhibiting cyclin D-dependent kinases CDK4 and CDK6.

of the gene. The enrichment at the 3′ terminus might be explained by the lack of functional domains downstream of the dual-coding region, obviating the need for a downstream alternative splicing event to restore the original reading frame. The apparent abundance of dual-coding regions within C termini may facilitate the preservation of partial functionality provided by single-coding upstream N-terminal domains.

Dual-Coding Exons Follow the Selective Regime of INK4a/ARF. As we showed previously, the two reading frames of the *INK4a/ARF* locus evolve asymmetrically, with ARF accumulating the bulk of amino acid substitutions. Our set of 90 dual-coding exons closely follows this trend. To study the evolutionary regime of dual-coding exons, we first compared the rates of nonsynonymous substitutions between constitutive (dN^{const}) and alternative (dN^{alt}) reading frames of each dual-coding exon (Fig. 5*A*). If the two reading frames were under similar selective pressure, we would expect dN^{const} to equal dN^{alt} as was previously reported for *GNAS1* and *XBP1* (20, 21). On the contrary, most dual-coding exons showed strongly asymmetrical evolutionary pressures acting on the two reading frames: constitutive frames, on average, evolved considerably slower than the alternative frames (Fig. 5*A*). This closely mimics the behavior of the

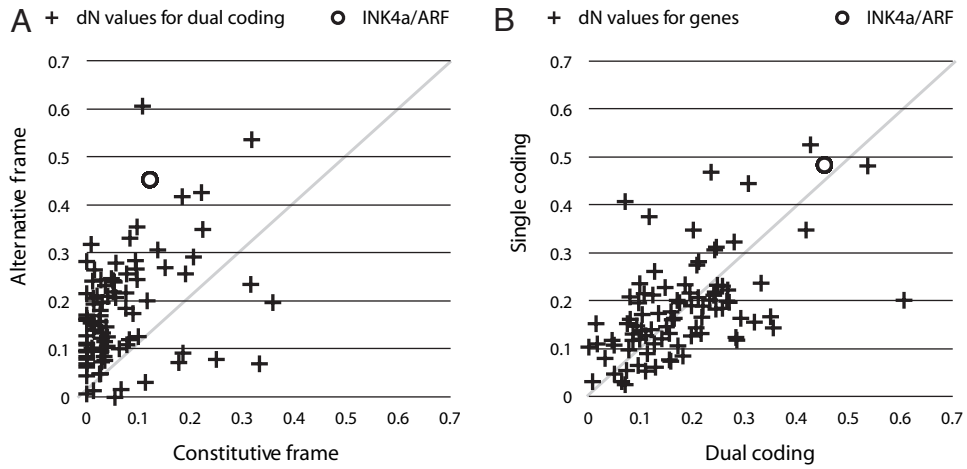

Fig. 5. Substitution dynamics of dual-coding exons. (*A*) Nonsynonymous substitutions in 90 dual-coding regions, plotted for the two ORFs of each region. A strong asymmetry between the rates of evolution in constitutive and alternative frames can be observed. (*B*) Nonsynonymous substitutions in the alternative frame, plotted for the alternative reading frame of each dual-coding region against the alternative frame of a single-coding region from the same gene.

reading frames within the *INK4a/ARF* locus ($dN^{const} = 0.12$ and $dN^{alt} = 0.45$; Fig. 5*A*). Next, we contrasted substitution rates between single and dual-coding regions. To do so, we first partitioned each transcript into dual-coding and single-coding portions (Fig. 1*B*) and estimated dN in the alternative reading frame for both partitions ($^{SC}dN^{alt}$ and $^{DC}dN^{alt}$ for single- and dual-coding portions, respectively). The artificial alternative reading frame for the single-coding region is never translated *in vivo* and has no biological function, but can provide a useful estimate of the background rate of nonsynonymous substitutions in the alternative frame of the gene ($^{SC}dN^{alt}$). On the other hand, $^{DC}dN^{alt}$, which was computed from a real dual-coding portion, is biologically meaningful, and, therefore, we expected a signature of purifying selection in the alternative frame of the dual-coding region to be manifested as $^{DC}dN^{alt} \ll {}^{SC}dN^{alt}$. To our surprise, $^{SC}dN^{alt}$ and $^{DC}dN^{alt}$ seemed to be strongly correlated (Fig. 5*B*). Thus, nonsynonymous changes in most existing alternative frames occur at the background rate. This may imply that the majority of alternative frames are either nonfunctional or, rather unlikely, under positive selection. Yet the *INK4a/ARF* locus, where both reading frames have been confirmed to be biologically relevant, behaves similarly, with effectively indistinguishable $^{DC}dN^{alt}$ and $^{SC}dN^{alt}$ (0.45 and 0.48, respectively; Fig. 5*B*).

The INK4a/ARF Paradox. The two reading frames of the *INK4a/ARF* locus control two essential tumor suppressor pathways but, at the same time, evolve rapidly and asymmetrically. This asymmetry may be a general property of overlapping reading frames. In addition to *INK4a/ARF*, it has been documented for the 90 dual-coding exons identified by our analysis of human and mouse splicing graphs, an independent survey of human genes (30), as well as in bacterial and viral systems (23, 32–34). The existing knowledge about the functions of the *INK4a/ARF* locus conflicts with this evolutionary trajectory. Although the INK4a frame evolves along a traditional path guided by purifying selection, it changes much faster than the majority of single coding genes (Fig. 3) including the remaining members in the INK4a/CDK4/CDK6/Rb pathway (Fig. 4*B*). The *ARF* frame evolves in the opposite direction: it accumulates nonsynonymous substitutions at an unusually high rate, yet it encodes a functional protein involved in a tumor suppression pathway. How can one explain this puzzling pattern? The obvious hypothesis that *ARF*-encoded protein has no biological function is contradicted by the overwhelming amount of experimental evidence. The locus behavior is consistent with this hypothesis: *INK4a* is under purifying selection ($dN < dS$) and *ARF* is under diversifying selection ($dN > dS$), whereas an unusually high dS within *INK4a* is likely a byproduct of its codependency with the rapidly changing *ARF*. There are documented cases of bacterial and viral overlapping reading frames that conform to this explanation (23). However, it is unclear whether the locus is in an evolutionary steady state, and if not, then what state it may be evolving toward. If *ARF* is under diversifying selection, then what is the moving target? If it is under directional selection, then what constitutes the ultimate fitness maximum for a tumor suppressor? A considerable uncertainty surrounds the role of selection in the evolution of cancer, making it difficult to address these questions (35–37). Finally, *ARF* may still be under purifying selection, and the observed pattern is due to relaxed constraints on the protein sequence: changes are allowed so long as they do not impair the codependent INK4a and fundamental physical properties, such as net charge, are kept within a permissible range. A number of observations argue in favor of this hypothesis. ARF domains responsible for binding to MDM2 are arginine rich (Arg is encoded by six synonymous codons), and their exact location and sequence seem to be less critical than the charge itself (38). The section of the *ARF* frame overlapping with *INK4a* frame (within exon 2) contains significantly ($P \ll 0$, Fisher's exact test) more synonymous sites than a single coding section of the same locus suggesting that it has evolved to sustain high substitution rate by highly redundant codons.

The benefit of dual coding in the *INK4a/ARF* locus is enigmatic. In two other well studied mammalian dual-coding genes, *GNAS1* and *XBP1*, overlapping reading frames are translated from a single mRNA: in *GNAS1*, this switch occurs by means of the use of alternative start codons (3); in *XBP1*, a specific endonuclease, IRE1, removes a short spacer (not in multiples of three) rerouting translation between reading frames (39). A potential advantage of containing overlapping reading frames in a single transcript is transcriptional and translational coupling (40). For example, a transcript of GNAS1 carrying overlapping reading frames is expressed in a small subset of cell types (41). Upon translation, the product of one reading frame, ALEX, specifically binds to the product of the other frame, XLαs, possibly regulating the signal transduction properties of the latter (3, 42). None of these features apply to *INK4a/ARF*: each of the frames is translated from its own transcript whose expression is not tightly coordinated. Yet one possible benefit of tight coupling between INK4a and ARF may be in the sharing of the 3′-UTR region. The two mRNAs exhibit extraordinary stability, which is thought to be determined primarily by 3′-end sequence elements (43). Another possibility is that the region of overlap may be sustained through evolution to promote variability to eliminate mutants (44). Whereas fragments of the gene (e.g., single coding exons) require high variability, still other fragments may be selected for lower amino acid changing substitutions. In this case, a single mutation in the region of overlap would lead to multiple amino acid substitutions, subsequently invoking an organism's surveillance mechanisms.

Our main finding is the apparent incompatibility of *INK4A/ARF* evolutionary dynamics with its function. The main purpose of a tumor suppressor is to provide reliable monitoring of the cell proliferation process. If this function is compromised, uncontrolled proliferation may begin leading to the development of cancer. Therefore, from the selection standpoint, it would make most sense to maintain a virtually invariable tumor suppressor. The mutation rate observed within *INK4a/ARF* draws a completely different picture, and the locus is mutated in more than half of known human cancers. Furthermore, the actual mutation rate occurring within each lineage of the mammalian tree may be considerably higher than the estimated substitution rate. In addition, our approach considers only changes that became fixed in respective populations and ignores somatic mutations. But the stability of tumor suppression with respect to somatic mutations is fundamentally important. For example, *INK4a/ARF* is inactivated in a number of human colorectal cancers. The inner lining of the human colon is an epithelial tissue where $\approx 10^{10}$ cells are replaced daily (45). This enormous turnover is ensured by continuous division of millions of stem cells, creating possibilities for somatic mutations. Combining these data with the already high intrinsic mutation rate of the *INK4a/ARF* locus makes it a liability for an organism.

Materials and Methods

Analysis of Nucleotide Substitutions. The MG94×HKY85 codon model with transition/transversion bias correction and nine base frequency parameters was fitted to sequence alignments by using the HyPhy package (46). For every branch in a phylogenetic tree, the expected number of synonymous substitutions per synonymous site (dS) and its counterpart for nonsynonymous substitutions (dN) were estimated by using maximum likelihood (47). When comparing dS between two genes (gene 1 and gene 2) based on two sequences, we evaluated the null hypothesis of $dS1 = dS2$ versus the alternative of $dS1 \geq dS2$, using a likelihood ratio test, with the asymptotic distribution of $(\chi_0^2 + \chi_1^2)/2$ (48). An analogous test was carried out for comparing dN1 and dN2. To correct for multiple testing, we used the false discovery rate controlling method of Benjamini and Hochberg (49). This method bounds the expected proportion of false positives out of all significant results at the

specified value, guaranteeing, for example that no more that 5/100 significant results are expected to be false positives for $P = 0.05$. To compare the substitution rate of the *INK4a/ARF* locus against the genome-wide average, we used a set of 8,467 human/mouse alignments of strictly orthologous protein-coding regions, which can be accessed at www.bx.psu.edu/~anton/share/hs_mm.align.txt.

RB and p51 Pathway Analysis. We applied a series of relative ratio tests to compare the rates of synonymous and nonsynonymous substitution among members of the RB and p53 tumor suppressor pathways. Three codon models, with a separate synonymous and nonsynonymous rate on each branch, were fitted to each pair of genes belonging to the same pathway. The unrestricted model (HA) did not impose any constraints on the rates; the relative ratio model (HR) constrained either synonymous or nonsynonymous rates to be proportional (with scaling factor R) among corresponding branches of two gene trees; the equal rate model (H0) further set the scaling factor to one, hence enforcing equal rates. For each pair, we first tested HR vs. HA, using a likelihood ratio with B − 1 (B is the number of tree branches) degrees of freedom and $P = 0.05$ (corrected for multiple tests, using the conservative Bonferroni approach). When HR could not be rejected, we further tested H0 vs. HR using a likelihood ratio test with 1 degree of freedom and $P = 0.05$ (Bonferroni corrected). When HA was chosen, the genes were classified as evolving differently among lineages; HR led to the classification of concordant evolution, which was accelerated in one of the genes; finally H0 was selected, the genes were evolving concordantly with statistically indistinguishable rates.

Finding Dual-coding Exons. A splicing graph for a gene, read from the altGraphX database represents all possible connections among exons of that gene (31) (SI Fig. 7*A*). We generated all possible translations for each splicing graph by traversing it from the start codon at the 5′ end (SI Fig. 7*B*). For example, a graph shown in SI Fig. 7 represents a gene with two transcripts and yields two distinct polypeptides. After translation of each graph, we looked for exons where a nucleotide could be, say, a first codon position in one path through the graph but a third codon position in another path. Knowing this information, we selected exons with this dual-coding property (SI Fig. 7*C*). This approach can detect dual-coding regions that span multiple exons. First, we identified for each dual-coding exon which of the two reading frames is constitutive (e.g., performing the main functions of the gene) and alternative (e.g., tissue-specific). For this purpose, we translated both reading frames of each dual-coding exon and compared these translations against protein sequences of all known human genes from the Ensembl database. In most cases (90%), this analysis gave an unambiguous assignment to constitutive and alternative frames. Second, because the overlap between *INK4a* and *ARF* reading frames is conserved in many mammals studied to date, we expected true dual-coding exons to be conserved as well. We used a di-tag analysis procedure (SI Fig. 7 *D–F*) to inspect the conservation of the splicing patterns. Our goal was to determine which exon combinations appearing in transcripts were conserved, i.e., more likely to be biologically relevant. To that end, we required that the two alternative splices at the beginning of a dual-coding exon in our data set were conserved in another mammal. Specifically, we extracted 20 nucleotides of exonic sequence flanking the splice site of dual-coding exons (tags; see SI Fig. 7*D*). Using pairwise mammalian genomic alignments (50), we obtained the orthologous sequences in another genome (e.g., mouse; SI Fig. 7*E*). We treated two concatenated tags (therefore di-tag analysis) as a signature of a splice and searched for the presence of the 40-nt di-tag in the transcript database of the other species. For that purpose, we used UniGene, a database of all known transcripts (see SI Table 3). Next, we checked for the presence of stop codons, in both reading frames, within 100 bp downstream of the splice site, discarding potential dual-coding exons with stop codons in either species. Because chance conservation of two, relatively long overlapping ORFs between two species that diverged millions of years ago is highly improbable, dual-coding regions identified by this procedure are likely to have biological function. Our final data set included dual-coding exons conserved in human and in at least one of four other mammalian species (mouse, rat, dog, or cow).

We thank Norman E. Sharpless for providing critical feedback. Comments by David Krakauer and anonymous reviewers were instrumental in interpreting the evolutionary behavior of the locus. Special thanks to Wen-Yu Chung, Samir Wadhawan, Bouke de Vries, and Erik Axel Nielsen for providing analysis tools. We also thank Izabela Makalowska and Wojciech Makalowski for insightful discussions. A.N. is supported by a Beckman Young Investigator Award, National Science Foundation DBI Grant 0543285, the Huck Institutes for the Life Sciences, and Eberly College of Science at Pennsylvania State University. S.K.P. is supported by National Institutes of Health Grants AI43638, AI47745, AI57167, and R01-GM66276, University of California Universitywide AIDS Research Program Grant IS 02-SD-701, and University of California, San Diego Center for AIDS Research/National Institute of Allergy and Infectious Diseases Developmental Award AI36214.

1. Kim WY, Sharpless NE (2006) *Cell* 127:265–275.
2. Quelle DE, Zindy F, Ashmun RA, Sherr CJ (1995) *Cell* 83:993–1000.
3. Klemke M, Kehlenbach RH, Huttner WB (2001) *EMBO J* 20:3849–3860.
4. Kozak M (2001) *EMBO Rep* 2:768–769.
5. Yoshida H, Matsui T, Yamamoto A, Okada T, Mori K (2001) *Cell* 107:881–891.
6. Sherr CJ, Weber JD (2000) *Curr Opin Genet Dev* 10:94–99.
7. Sherr CJ (2001) *Nat Rev* 2:731–737.
8. Sharpless NE (2005) *Mutat Res* 576:22–38.
9. Bertwistle D, Sugimoto M, Sherr CJ (2004) *Mol Cell Biol* 24:985–996.
10. Lindstrom MS, Zhang Y (2006) *Cell Biochem Biophys* 46:79–90.
11. Janzen V, Forkert R, Fleming HE, Saito Y, Waring MT, Dombkowski DM, Cheng T, DePinho RA, Sharpless NE, Scadden DT (2006) *Nature* 443:421–426.
12. Krishnamurthy J, Ramsey MR, Ligon KL, Torrice C, Koh A, Bonner-Weir S, Sharpless NE (2006) *Nature* 443:453–457.
13. Molofsky AV, Slutsky SG, Joseph NM, He S, Pardal R, Krishnamurthy J, Sharpless NE, Morrison SJ (2006) *Nature* 443:448–452.
14. Robertson KD, Jones PA (1999) *Oncogene* 18:3810–3820.
15. Ruas M, Peters G (1998) *Biochim Biophys Acta* 1378:F115–177.
16. Yang G, Rajadurai A, Tsao H (2005) *J Invest Dermatol* 125:1242–1251.
17. Pomerantz J, Schreiber-Agus N, Liegeois NJ, Silverman A, Alland L, Chin L, Potes J, Chen K, Orlow I, Lee HW, *et al.* (1998) *Cell* 92:713–723.
18. Kim SH, Mitchell M, Fujii H, Llanos S, Peters G (2003) *Proc Natl Acad Sci USA* 100:211–216.
19. Zhang Y, Xiong Y (1999) *Mol Cell* 3:579–591.
20. Nekrutenko A, He J (2006) *Trends Genet* 22:645–648.
21. Nekrutenko A, Wadhawan S, Goetting-Minesky P, Makova KD (2005) *PLoS Genet* 1:e18.
22. Holmes EC, Lipman DJ, Zamarin D, Yewdell JW (2006) *Science* 313:1573.
23. Rogozin IB, Spiridonov AN, Sorokin AV, Wolf YI, Jordan IK, Tatusov RL, Koonin EV (2002) *Trends Genet* 18:228–232.
24. Gilley J, Fried M (2001) *Oncogene* 20:7447–7452.
25. Pazos F, Valencia A (2001) *Protein Eng* 14:609–614.
26. Goh CS, Cohen FE (2002) *J Mol Biol* 324:177–192.
27. Mintseris J, Weng Z (2005) *Proc Natl Acad Sci USA* 102:10930–10935.
28. Muse SV, Gaut BS (1997) *Genetics* 146:393–399.
29. Ihmels J, Collins SR, Schuldiner M, Krogan NJ, Weissman JS (2007) *Mol Syst Biol* 3:86.
30. Liang H, Landweber LF (2006) *Genome Res* 16:190–196.
31. Sugnet CW, Kent WJ, Ares M, Jr, Haussler D (2004) *Pac Symp Biocomput* 9:66–77.
32. Jordan IK, Sutter BA, IV, McClure MA (2000) *Mol Biol Evol* 17:75–86.
33. Hughes AL, Hughes MA (2005) *Virus Res* 113:81–88.
34. Pavesi A (2006) *J Gen Virol* 87:1013–1017.
35. Merlo LM, Pepper JW, Reid BJ, Maley CC (2006) *Nat Rev Cancer* 6:924–935.
36. Crespi BJ, Summers K (2006) *Biol Rev Cambridge Philos Soc* 81:407–424.
37. Thomas MA, Weston B, Joseph M, Wu W, Nekrutenko A, Tonellato PJ (2003) *Mol Biol Evol* 20:964–968.
38. Bothner B, Lewis WS, DiGiammarino EL, Weber JD, Bothner SJ, Kriwacki RW (2001) *J Mol Biol* 314:263–277.
39. Mori K (2003) *Traffic* 4:519–528.
40. Krakauer DC (2000) *Evol Int J Org Evol* 54:731–739.
41. Plagge A, Gordon E, Dean W, Boiani R, Cinti S, Peters J, Kelsey G (2004) *Nat Genet* 36:818–826.
42. Freson K, Jaeken J, Van Helvoirt M, de Zegher F, Wittevrongel C, Thys C, Hoylaerts MF, Vermylen J, Van Geet C (2003) *Hum Mol Genet* 12:1121–1130.
43. Hara E, Smith R, Parry D, Tahara H, Stone S, Peters G (1996) *Mol Cell Biol* 16:859–867.
44. Krakauer DC, Plotkin JB (2002) *Proc Natl Acad Sci USA* 99:1405–1409.
45. Potten CS, Booth C, Pritchard DM (1997) *Int J Exp Pathol* 78:219–243.
46. Pond SL, Frost SD, Muse SV (2005) *Bioinformatics* 21:676–679.
47. Muse SV (1996) *Mol Biol Evol* 13:105–114.
48. Self SG, Liang KY (1987) *J Am Stat Assoc* 82:605–610.
49. Benjamini Y, Hochberg Y (1995) *J R Stat Soc B Met* 57:289–300.
50. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W (2003) *Genome Res* 13:103–107.