6514–6530 *Nucleic Acids Research, 2013, Vol. 41, No. 13* *Published online 6 May 2013*
doi:10.1093/nar/gkt274

Identification of the nature of reading frame transitions observed in prokaryotic genomes

Ivan Antonov[1], Arthur Coakley[2], John F. Atkins[2], Pavel V. Baranov[2] and Mark Borodovsky[1,3,4,5,*]

[1]School of Computational Science and Engineering at Georgia Tech, Atlanta, GA 30332, USA, [2]Department of Biochemistry, University College Cork, Ireland, [3]Department of Biological and Medical Physics, Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region 141700, Russia, [4]Center for Bioinformatics and Computational Genomics at Georgia Tech and [5]Joint Georgia Tech and Emory Wallace H Coulter Department of Biomedical Engineering, Atlanta, GA 30332, USA

Received November 12, 2012; Revised February 22, 2013; Accepted March 22, 2013

ABSTRACT

Our goal was to identify evolutionary conserved frame transitions in protein coding regions and to uncover an underlying functional role of these structural aberrations. We used the *ab initio* frameshift prediction program, GeneTack, to detect reading frame transitions in 206 991 genes (fs-genes) from 1106 complete prokaryotic genomes. We grouped 102 731 fs-genes into 19 430 clusters based on sequence similarity between protein products (fs-proteins) as well as conservation of predicted position of the frameshift and its direction. We identified 4010 pseudogene clusters and 146 clusters of fs-genes apparently using *recoding* (local deviation from using standard genetic code) due to possessing specific sequence motifs near frameshift positions. Particularly interesting was finding of a novel type of organization of the *dnaX* gene, where recoding is required for synthesis of the longer subunit, τ. We selected 20 clusters of predicted *recoding* candidates and designed a series of genetic constructs with a reporter gene or affinity tag whose expression would require a frameshift event. Expression of the constructs in *Escherichia coli* demonstrated enrichment of the set of candidates with sequences that trigger genuine programmed ribosomal frameshifting; we have experimentally confirmed four new families of programmed frameshifts.

INTRODUCTION

Protein encoding imposes constraints on genomic sequence. Because the constraints are frame dependent it is possible to infer from a genomic sequence, which one out of six possible reading frames is likely to be translated (if any). Recently, we have developed a computational method, GeneTack, for identifying such infrequent locations where protein coding instantly transits from one frame to another without presence of stop and start codons (1). In the present work, we use comparative genomics to classify frame transitions predicted by the new method in prokaryotic genomes. This approach is conceptually similar to one recently used in a study of bacterial genes annotated in GenBank as genes with disrupted open reading frames (ORFs) (2).

There are several reasons, both technological and biological, to observe frame transitions in prokaryotic genomes. On the technological side, sequencing and assembly errors produce artifacts subsequently incorporated into databases. Biological reasons are indel mutations, conserved in evolution frame transitions involved in non-standard mechanisms of transcription or translation known as *recoding* (3–8), phase variation, etc.

Many indel mutations that may truncate and inactivate protein products would not affect the rest of the sequence, particularly the promoter region. Therefore, a mutated gene (a pseudogene) may still be transcribed with the RNA potentially carrying a regulatory role; thus, in a certain lineage the pseudogene sequence may evolve under purifying selection. Frame transitions also appear in genes that use phase variation, e.g. at a specific hypermutable location (9,10).

Identification of genes with *recoding* from genomic sequence alone is a challenging task (11–13). New *recoding* events may reveal novel DNA sequence patterns (stimulatory sequences) required for switching to a non-standard mechanism of gene expression. Understanding the ways stimulatory sequences work can shed light on yet unknown details of transcription and translation machinery. It may also provide new synthetic

*To whom correspondence should be addressed. Tel: +1 404 894 8432; Fax: +1 404 894 4243; Email: borodovsky@gatech.edu

biology means for controlling gene expression. Therefore, this study has a particular emphasis on identification of novel *recoding* candidates.

Recoding can work through a range of RNA editing mechanisms [slippage (14,15), guided RNA editing (16,17), adenosine and cytosine deamination (18–20)]. On the other hand, RNA transcripts may be subjects for a variety of translational *recoding* mechanisms [ribosomal frameshifting, codon redefinition, translational bypass, StopGo (21)]. Here we concentrate only on the mechanisms related to transitions between reading frames: ribosomal frameshifting and transcriptional realignment; given that these mechanisms have functional roles, they are often described as programmed, e.g. Programmed Ribosomal Frameshifting (PRF) and Programmed Transcriptional Realignment (PTR) (2).

On PRF a ribosome changes the initial reading frame at a specific location in mRNA. Displacements of a ribosome by +1 and −1 nucleotide have been predominant while displacements by +2, −2 and even up to 50 nucleotides (commonly known as bypassing) have been documented (22–26). While PRF has been detected in many prokaryotic and eukaryotic species, it is especially prevalent in viruses (5,27). High-efficiency PRF is modulated by a range of stimulatory signals at the RNA level (28,29). Other signals can also affect readout of mRNA either through complementary mRNA:rRNA pairing (30,31) or via nascent peptide interaction with the peptide exit tunnel of the ribosome (32,33).

The PTR event [also termed transcriptional slippage (15), stuttering (34), molecular misreading (35) and reiterative transcription (36)] occurs when realignment of a growing RNA chain to the DNA template within the RNA polymerase ternary complex results in insertion or deletion of a single or multiple nucleotides relative to the DNA template (37,38). The indels usually occur in characteristic motifs such as homopolymeric runs of adenines or thymines.

In prokaryotes, the best known examples of genes with *recoding* are Insertion Sequence (IS) elements, as well as genes for Release Factor 2 (*prfB*) and DNA polymerase III (*dnaX*). Over 80% of eubacterial species use +1 PRF during *prfB* expression (11,39). Fewer examples of *recoding* were reported for *dnaX* (7,40) whose expression in diverse organisms has been less studied. Interestingly, though in different species, both PTR and PRF are known to be used in expressing *dnaX* (40), suggesting that PTR and PRF, at least in some locations, are interchangeable. IS elements, in particular members of IS3 family, use both PRF and PTR for expressing protein products (2). *Recoding* mechanisms are especially abundant among viruses (5,41,42). In this study, we sought to identify previously unknown cases of *recoding* that involve frame transitions. Our bioinformatic approach is independent from prior annotation; this approach yielded a number of new *recoding* candidates with some subjected to experimental tests.

MATERIALS AND METHODS

Translation of predicted fs-genes; BLASTp and Pfam validations

A frame transition predicted by GeneTack (1) indicates the presence of two overlapping protein coding ORFs that may constitute a single gene with a frameshift, an 'fs-gene'. Frequently, these two ORFs are annotated in databases as separate genes. For further analysis, we combine the first ORF with the second ORF at the position of the predicted frameshift taking into account the predicted frameshift direction (+1 or −1). Translation of the ORF extended in this manner yields an 'fs-protein'. Thus, each predicted frameshift makes an fs-gene and a fs-protein.

The GeneTack false discovery rate (FDR) determined earlier on a set of 17 prokaryotic genomes is ~32% (1). Given the relatively high FDR, we used two complementary methods to confirm GeneTack predictions. We used BLASTp search to find a protein in the NCBI nr database whose alignment to the fs-protein (the query) had a score with E-value $<10^{-10}$. Moreover, the sequence alignment to database protein had to cover at least a 100 AA fragment of the fs-protein containing the predicted frameshift position (Supplementary Figure S1A). If, on the other hand, the BLASTp search for an fs-protein query produced two sets of BLASTp hits disconnected at the frameshift position (Supplementary Figure S1B), this result was indicative of the presence of a pair of overlapping genes. Then, the predicted frameshift was characterized as an instance of frame transition between two adjacent genes. Still, given the possibilities of gene fusion and fission, some of BLASTp validated frameshifts may yet be false positives, e.g. a frameshift predicted between adjacent genes whose homologs are fused in another genome (43).

The frameshift validation could be more substantial were a search against the Pfam domain database produced an alignment to a Pfam domain (with E-value $<10^{-3}$) covering the predicted frameshift position. Assuming that a conserved domain could not be divided between two fused genes, the Pfam confirmation would exclude artifacts related to gene fission and fusion.

Ribosome-binding site of the downstream protein-coding region

A ribosome-binding site (RBS) motifs are not expected to be specifically associated with frameshifts caused by indel (pseudogenizaton) mutations, as well as by sequencing error. However, some programmed frameshifts have stimulatory sequences of the Shine-Dalgarno type located near the start of the protein-coding part of the downstream ORF. The gene prediction program GeneMarkS (44) provides parameters and initial gene predictions for GeneTack as well as computes an RBS score for the upstream region of each predicted gene; in our observations the RBS score values range between −11 and 8 (a larger score corresponds to a stronger RBS). Normally, in a location of given fs-gene, GeneMarkS predicts two separate genes. If the downstream 'gene' is

not a real full-length gene, its 'RBS score' is expected to be low. Therefore, GeneTack filters out fs-genes if the RBS scores for downstream coding region is >2.2 (1). Still, in analysis of clusters of fs-genes, even slightly elevated RBS scores that appear consistently in the whole cluster could be indicators of false-positive predictions. For 10 434 frameshifts confirmed by both BLASTp and Pfam, an average value of the RBS scores of downstream coding regions was −1, while the average value of the RBS scores for downstream coding regions in all the remaining fs-genes was −0.14.

Clustering

All 206 991 predicted fs-proteins (with or without BLASTp and Pfam confirmations) were grouped into clusters based on sequence similarity and conservation of frameshift position and directionality (+1 or −1). First, in the database of all fs-proteins 'all-against-all' BLASTp search was performed with a stringent E-value threshold 10^{-50} chosen to avoid inclusion of non-homologous proteins in the clusters that would facilitate detection of conserved DNA motifs related to programmed frameshifts.

Next, a graph was built with nodes representing 206 991 fs-proteins. Two nodes were connected by an edge if (i) positions of two frameshifts were inside the BLASTp pairwise alignment block (separated from the block border by at least 10AA); (ii) both frameshifts had the same direction (+1 or −1); and (iii) the distance between frameshift positions in the block was ≤50AA. This graph-connected components with two or more nodes were called clusters (GeneTack clusters).

The fs-genes that did not cluster were likely to be related to genes sequenced with errors, orphan pseudogenes, pairs of overlapping orphan genes or even orphan genes with programmed frameshifts.

Then we proceeded with classification of clusters of homologous fs-genes as (i) fs-genes with programmed frameshifts; (ii) pseudogenes or hypothetical pseudogenes; (iii) fs-genes with phase variation; (iv) fs-genes with translational coupling; as well as (v) fs-genes related to overlapping pairs of homologous genes (false-positive clusters).

Functional characterization of the GeneTack clusters

In a given GeneTack cluster, we expected to see fs-proteins with similar function. If more than 50% of fs-proteins in a cluster contained the same Pfam domain, its name was assigned to the cluster. Clusters of multi-domain proteins received a 'multi-function name' with more frequent domains listed first.

If the majority of fs-proteins in a cluster did not have a match to a Pfam domain, the cluster did not receive a Pfam-derived name (just a cluster ID), unless a functional cluster name was derived by BLASTp 'function transfer' from hits to the NCBI nr database produced by the fs-proteins.

Identification of clusters of fs-genes using non-standard mechanisms of transcription and translation

Transcriptional realignment and ribosomal frameshifting often occur at specific sequences where the PTR or PRF efficiency is augmented by additional *cis*-elements (Figure 1). Owing to the limited repertoire of shift- and slip-prone sequences, they evolve under purifying selection. In case of PTR, the specific sequences often appear to be homopolymers of eight nucleotides or longer, or combinations of two shorter homoploymers.

A PRF event involves rearrangements of two tRNAs interacting with two codons. For a +1 frameshift, the tRNA recognizing the A-site moves one nucleotide downstream, while in the case of a −1 frameshift, both tRNAs already occupying P- and A-site move one nucleotide upstream. Thus, in both cases, there are seven nucleotides involved in the PRF frameshift mechanism (45). On the other hand, the PTR slippage sites have been observed to be even longer (38). Therefore, it is expected that genuine instances of programmed frameshifting should be related to at least seven-nucleotide-long sequences evolving under purifying selection. Identification of conserved hepameric and longer sequences at the predicted frameshift site can be used as supportive evidence for programmed frameshifting.

To precisely delineate specific motifs related to PTR or PRF in a given cluster, we built a multiple alignment of 'frameshift boxes', sequences surrounding predicted frameshift positions. A frameshift box is bounded by two stop codons, one at the 5′-end of the downstream ORF and the other at the 3′-end of the upstream ORF (Figure 2). Both predicted and true frameshift positions should have occurred within the frameshift box. If the distance between the two stop codons was >100 nt (a frequent case in high GC genomes), the frameshift box was reduced to the 100-nt long vicinity of the predicted frameshift.

Several efficient algorithms and software tools for finding conserved motifs have been developed earlier [e.g. the Gibbs Sampler (GS) (46) and MEME (47)]. Still, in the case of PRF not only the motif *per se* but also the phase of motif with respect to reading frame, set by the start codon of upstream ORF, is important. For example, in the *prfB* gene encoding Release Factor 2, the consensus frameshift motif is YTT_TRA_C, with the triplet TRA being a stop codon. When phase is important (e.g. PRF motifs) the DNA sequence alphabet could be extended by an additional symbol indicating the frame of upstream ORF (an underscore symbol). Frameshift box sequences phased by underscores were used in a customized version of the GS algorithm to produce motifs with a given triplet phase. The consensus of motif sequences (a phased motif) was used to characterize the frameshift site in a given cluster.

To initially identify motifs prone to +1 and −1 programmed frameshifts, we searched for framed heptamers that occured in the frameshift boxes of a given cluster (N_NNN_NNN for −1 frameshifts and NNN_NNN_N for +1). Clusters containing between 5

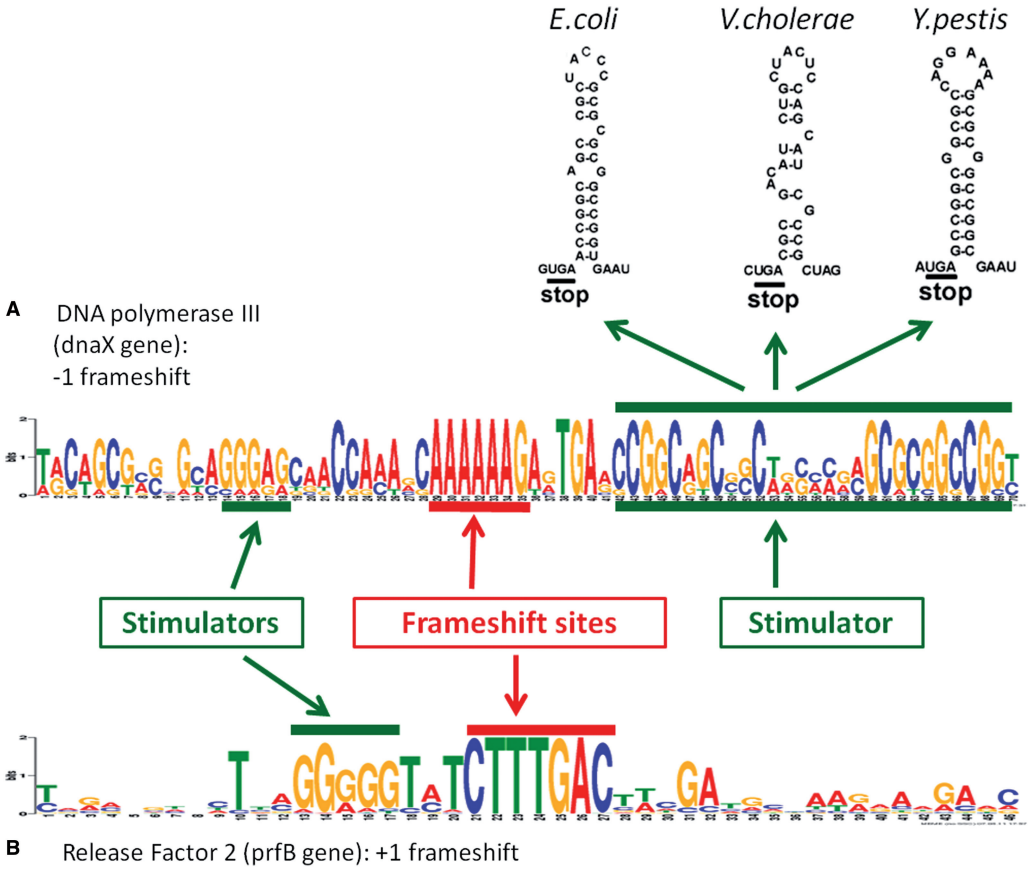


Figure 1. Examples of patterns facilitating programmed frameshifting. (**A**) '−1' programmed frameshift is used in *dnaX* gene to express two subunits of DNA polymerase III. The Logo for (A) was derived from aligned sequences from 9 genera (*Escherichia*, *Salmonella*, *Neisseria*, *Vibrio*, *Shigella*, *Citrobacter*, *Enterobacter*, *Yersinia*, *Serratia*). The frameshift signal consists of conserved frameshift pattern AAA_AAA_G ('slippery sequence') and two stimulators. The upstream stimulator is a Shine-Dalgarno–like sequence that interacts with ribosome, while the downstream stimulator makes a hairpin secondary structure (7). (**B**) '+1' programmed frameshift is used in *prfB* gene to auto regulate expression of Release Factor 2. The Logo for (B) was derived from 413 sequences (138 genera). The frameshift signal consists of conserved frameshift motif with consensus CTT_TGA_C and upstream Shine-Dalgarno–type sequence stimulator.

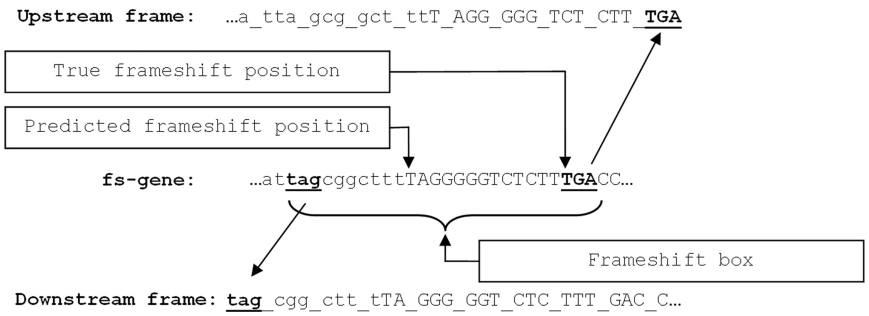


Figure 2. An example of a 'frameshift box'. Predicted frameshift position appears in between two stop codons situated in different frames (TAG stop codon upstream and the TGA stop codon downstream). The true frameshift position is always located inside the 'frameshift box', the region between two stop codons.

and 100 fs-genes with average sequence identity of the frameshift box ≤80% were selected (1017 '−1' clusters and 1380 '+1' clusters). Starting positions of motifs were chosen randomly and GS was run 100 times searching for N_NNN_NNN motifs in −1 clusters and NNN_NNN_N motifs in +1 clusters.

For large clusters (with 100 or more fs-genes), alignments of the most over-represented heptamers were used to initiate the GS iterations. Consensus sequences of alignments found by the GS were recorded (framed heptamers). When motif positions for the first iteration were chosen randomly, consensus heptamers found in

Table 1. Features (the first column) used to classify predicted frameshifts into types (the type names are given in the top two rows)

	Cluster type					Singleton type	
	Programmed frameshift	Phase Variation	Translational Coupling	Pseudogene	H-pseudogene	Pseudogene	H-pseudo / Error
Cluster contains 5 or more fs-genes	Yes	Yes	Yes	n/r	n/r	n/a	n/a
Conserved frameshift site	Yes	n/r	n/r	No	No	n/a	n/a
Cluster with small (≤2) number of genera	n/r	n/r	n/r	Yes	Yes	n/a	n/a
RefSeq annotation of a pseudogene[a]	n/r	n/r	n/r	Yes	No	Yes	No
Tandem repeat near frameshift position	n/r	Yes	n/r	n/r	n/r	n/r	n/r
ORF2 start is located close to ORF1 stop	n/r	n/r	Yes	n/r	n/r	n/r	n/r
BLASTp validation[b]	n/r	n/r	n/r	n/r	Yes	n/r	Yes
Pfam validation	n/r	n/r	n/r	n/r	n/r	n/r	Yes
Manual verification[c]	n/r	Yes	Yes	n/r	n/r	n/r	n/r

H-pseudo, hypothetical pseudogene; n/r, the feature is not required; n/a, the feature is not applicable; Error, sequencing error.
[a]A cluster must contain at least one annotated pseudogene.
[b]>50% of cluster fs-genes must be validated by BLASTp.
[c]Manual verification includes functional analysis of the fs-proteins and literature survey.

different GS run could vary. We recorded a number of times a heptamer X appeared as a GS consensus for a particular cluster. The score of a heptamer X was computed as follows:

$$Score(X) = \sum_{clusters} cluster\_size \bullet \frac{Number\ of\ times\ X\ is\ found}{Number\ of\ GS\ runs}$$

Notably, consensus heptamers containing a start codon for a downstream frame (AT_G for '+1' and A_TG for '−1') indicated that frameshifts were predicted at overlaps of pairs of homologous genes.

Among the consensus heptamers found in our analysis, there were seven A-rich heptamers (AAA_AAA_A, AAA_AAA_T, A_AAA_AAG, T_AAA_AAA, A_AAA_AAC, A_AAA_AAA and G_AAA_AAA) (Supplementary Table S1). It was reassuring to see this result, as it is well known that poly-A motifs frequently appear in frameshift sites.

To better identify possible stimulatory motifs, the alignments were extended 20 nt upstream and downstream from detected motifs of frameshift sites. We used positional nucleotide frequencies defined in extended alignments (with frame phase omitted) to build a logo (48) of sequence conservation at the frameshift site. Obviously, finding a conservation pattern did not guarantee that a given fs-gene cluster contained genes with programmed frameshifts. Evolutionary conserved sequences could be present at overlaps of homologous gene pairs. Therefore, we introduced several features for cluster classification as described in Table 1.

Inferring a type of frame transition mechanism

At poly-A/T sites, programmed frame transition may occur during either transcription or translation (for example in transposase and in *dnaX* genes). Sequence conservation features allow for selecting a specific hypothesis on the mechanism of programmed frameshift.

We have observed that in genes using PRF, the frameshift direction, +1 or −1, is conserved among orthologs (for example, all *prfB* genes use +1 shifts). On the other hand, our data showed that in genes with confimed PTR, such as transposase genes, the orthologs do not necessarily keep the same frameshift direction. Thus, for a single transposase family, our clustering approach produced two separate clusters, with +1 and −1 frameshifts. For example, the HTH_Tnp_IS630 family had 495 members in the '−1' cluster and 185 members in the '+1' cluster (Figure 3). In each cluster, only a framshift in specific direction could lead to synthesis of a full-length transposase.

Given these observations, if there were two or more clusters of fs-genes with the same function but different frameshift direction, the predicted nature of frameshift was PTR; otherwise, if only one direction of frameshift was ubiquitous in a cluster, the predicted mechanism was PRF.

Interestingly, in some experimentally confirmed genes with *recoding*, the mechanism of programmed frameshifting is still under debate as even orthologous genes may use different *recoding* mechansms. For example, expression of *dnaX* genes goes via PRF in some prokaryotic species and via PTR in other (40,49).

Frame transitions related to phase variation and translational coupling

Phase variation, a reversible and inheritable change of bacterial phenotype is often considered as a random process evolved to facilitate evasion of a host immune respond. Among molecular mechanisms of phase variation (homologous recombination, inversion of DNA elements, etc.), slipped strand mispairing (SSM) seems to be the most common. During replication, SSM may occur at repeat units (such as short sequence repeats, microsatellites or variable-number tandem repeats). The repeat unit could be as simple as a homopolymer sequence [e.g. poly-A in the *p78* gene of *Mycoplasma fermentans* (50) or poly-C/poly-G in the type III methyltransferases genes (51)] or a repeat of more complex subunits (for example AGTC is repeated >30 times in the *mod* gene of

Haemophilus influenzae). Insertion or deletion of a repeat unit on replication creates a frameshift mutation to turn the gene on and off.

Phase variation has been studied mainly in bacterial pathogens; however, it may occur in non-pathogens as well (52). The majority of proteins encoded by genes involved in phase variation are exposed to the cell environment. Examples include proteins involved in capsule, fimbriae, pili, flagella as well as surface proteins: transporters, receptors and porins. Notably, many of large GeneTack clusters contain fs-genes for cell surface and secretory proteins. Still, phase variation has also been associated with DNA modification and metabolism-associated genes (53).

As poly-AT is a slippery sequence for DNA polymerase (as well as for RNA polymerase and ribosomes), a stretch of poly-AT could cause phase variation. For a given cluster, we computed a fraction of fs-genes with a poly-AT stretch (minimal length 7 nt) close to the frameshift, designated as %AT. Finally, we used the tandem repeat finder program (54) to identify other types of repeats (such as poly-G or poly-GC). The program parameters were set to report homopolymers as repeats (minimal length 7 nt). For a given cluster, we determined a fraction of fs-genes with tandem repeats (other than Poly-A and poly-T) near a predicted frameshift, designated as %R.

Clusters of fs-genes were classified as related to phase variation if (i) we observed characteristic repeats near the frameshift position as well as (ii) function of some fs-genes in the cluster was earlier associated with phase variation.

Translational coupling of two adjacent genes implies existence of a re-initiation mechanism that requires proximity of the translation initiation site of downstream gene to the translation termination site of upstream gene. The downstream gene may have a weak RBS site. Gene pairs with evolutionary conserved translational coupling could be predicted as fs-genes and could form clusters. Observation of evolutionary conservation of co-location of upstream ORF stop and downstream ORF start codons would support a translational coupling hypothesis. For a cluster with $\geq$100 fs-genes, we determined a fraction (%S) of fs-genes with co-localized starts and stops (within 10 nt distance). A high value of %S was considered to be a signature of translational coupling. Phase variation or translational coupling characterization, as shown in Table 2, was based on the combination %S, %R and %AT values.

Experimental verification of predicted programmed frameshifting

Bacterial strains

The *E. coli* strains DH5α and MG1655Δ*lacIZ* were used for plasmid propagation and western blot analyses, respectively. Strains were grown in Luria–Bertani (LB) plus or minus isopropyl-β,D-thiogalactopyranoside (IPTG).

Plasmid construction

The vector pJ307 was derived from the GST-MBP-His fusion vector (pGMH57) by ligating annealed oligonucleotides (5′-GATCAGCTCGAGCACTAGTCC ATGGGGATCCAAG-3′ and 5′-AATTCTTGGATCCC CATGGACTAGTGCTCGAGCT-3′) into pGMH57 between BamHI-EcoRI restriction sites of pGHM57 (55). Twenty inserts were constructed by PCR amplification of complementary oligonucleotides to produce a full-length sequence containing 5′ XhoI and 3′ BglII restriction sites. These fragments were restriction digested and then ligated into the vector pJ307, digested by compatible restriction enzymes PspXI and BamHI (present in the new cloning site of pJ307), so that the *MBP* gene was in an alternative frame (+1 or −1) relative to *GST* or in-frame for positive control. Supplementary Table S2 shows the full-length sequences of the inserts.

Western blot analysis

Overnight cultures of strains expressing the appropriate plasmid were diluted 1:100 in LB Broth, grown for 2 h at 37°C, and then induced with 100 mM IPTG for an additional 2 h at 37°C. Crude extracts were obtained by culture centrifugation and re-suspending the bacterial pellet in Laemmli sample buffer. Proteins were separated on 10% sodium dodecyl sulphate polyacrylamide gel electrophoresis gels and transferred to nitrocellulose membranes (Protran). Immunoblots were incubated at 4°C overnight in 5% milk/phosphate buffered saline–Tween containing a 1:500 dilution of rabbit anti-GST or 1:2000 dilution of rabbit anti-HIS. Immunoreactive bands were detected on membranes after incubation with appropriate fluorescently labeled secondary antibodies using a LI-COR Odyssey® Infrared Imaging Scanner (LI-COR Biosciences). The amounts of termination and frameshift product were quantified by ImageQuant. The frameshifting efficiency was estimated as the ratio of the amount of frameshift product to the total amount of termination plus frameshift products.

RESULTS

Frame transitions predicted in 1106 genomes

We downloaded 1106 prokaryotic genomes longer than 1 Mb (77 archaeal and 1029 bacterial; see Table 3 for information on phylogenetic diversity) from the NCBI Web site ftp://ftp.ncbi.nih.gov/genomes/Bacteria/all.gbk.tar.gz (on 12 April 2010; draft genomes were excluded). The GeneTack-GM software program (1) with default settings was used to screen all the sequences; 206 991 frameshifts were predicted. The number of predicted frameshifts in a given genome has shown correlation with its length and the number of genes (see Supplementary Figure S2). As the GeneTack accuracy in frameshift detection is characterized by 32.8% FDR, we expected about one-third of the predictions to be related to frame transition between adjacent genes rather than to frameshifts. For translations of 36 668 (17.7%) fs-genes, BLASTp detected the NCBI nr database homologous proteins 'bridging the frameshifts'; also the Pfam domains covering predicted frameshifts were detected for 16 307 fs-genes (both continuous BLASTp hits and Pfam domains existed for 10 434 fs-genes). We have observed

Table 2. The largest clusters containing 100 or more fs-genes

Cluster ID	Cluster name	Size	#G	D	%AT	%R	%S	%B	BR
474411093	Release factor 2	428	138	+1	49	4	1	2	PF[a]
675840861	HTH_Tnp_1 (Transposase)	1699	106	−1	72	7	3	75	PF
188472814	DDE_Tnp_1 (Transposase)	384	37	−1	67	4	2	85	PF
888244788	DDE_Tnp_1 (Transposase)	108	5	+1	80	0	0	95	PF
667870043	HTH_Tnp_IS630 (Transposase)	495	20	−1	98	1	0	96	PF
858558073	HTH_Tnp_IS630 (Transposase)	185	28	+1	100	5	0	86	PF
696263973	DDE_Tnp_IS1 (Transposase)	230	8	−1	90	0	0	72	PF
784826247	Transposase IS911/IS222	112	5	−1	6	0	0	0	PF
752989859	Kinase/Phosphatase	105	23	+1	67	1	16	0	PF
279791230	HATPase_c, HisKA, Response_reg	594	148	+1	57	5	35	13	PV
487884579	HATPase_c, HisKA, Response_reg	292	98	+1	36	18	31	35	PV
107592512	HATPase_c, HisKA, Response_reg	162	51	−1	36	4	56	5	PV
437298609	BPD transporter	238	79	+1	34	9	34	5	PV
672517721	BPD transporter	149	46	+1	41	14	42	26	PV
953823467	BPD transporter	100	22	−1	5	17	10	0	PV
6376240	tRNA synthetase	215	81	+1	60	7	36	1	PV
138502135	Aminotransferase	175	88	+1	38	13	30	13	PV
354349696	Secretion system	140	51	+1	41	6	18	9	PV
322052632	Fucose synthase / Dehydratase	139	78	+1	44	9	37	4	PV
631171255	PqiA integral membrane protein	126	38	+1	71	4	17	5	PV
222950006	ABC transporter	436	116	+1	46	7	47	59	TC
785097185	ABC transporter	298	66	+1	62	4	48	63	TC
208900412	ABC transporter	293	97	+1	24	4	61	69	TC
624178257	ABC transporter	289	102	−1	49	8	45	64	TC
79330857	ABC transporter	280	97	+1	35	9	86	1	TC
104388297	ABC transporter	146	61	−1	21	18	64	11	TC
22890314	ABC transporter	144	49	+1	24	3	65	69	TC
471276212	ABC transporter	126	48	+1	25	2	60	33	TC
548076848	Flagella	139	34	+1	48	3	71	0	TC
585180489	Flagella	111	36	+1	8	11	75	0	TC
181132644	Flagella	118	46	+1	36	10	70^{2}	0	TC
847934252	Polyketide cyclase	132	38	+1	46	4	81^{2}	0	TC
697472870	Biotin carboxylase	128	44	+1	73	5	75	2	TC
876288400	Hydrolase/Epimerase	121	27	+1	28	2	79	0	TC
458305551	Polyphosphate kinase	113	35	+1	27	8	63	2	TC
237996460	Mur ligase	112	33	+1	78	3	$83^{1,2}$	90	TC
717516549	Epimerase	111	65	−1	44	13	61	2	TC
539781944	Oxidoreductase	109	51	−1	39	1	80	0	TC
515287573	Recombination factor RarA	104	52	+1	55	5	74^{2}	4	TC
984773919	Thymidylate kinase	100	25	+1	93	7	3	3	TC*

Size, number of fs-genes in the cluster; #G, number of different genera in the cluster; D, frameshift direction; %AT, fraction of fs-genes with 7+ nt poly-AT stretch located near predicted frameshift; %R, fraction of fs-genes with tandem repeats located near predicted frameshifts; %S, fraction of fs-genes with ORF2 start codon ATG (^{1}GTG) located within 10 nt (^{2}20 nt) from the ORF1 stop codon; %B, fraction of fs-proteins validated by BLASTp against NCBI nr database; BR, predicted biological role (PF, programmed frameshifting, PV, phase variation, TC, translational coupling).
[a]Experimentally verified.

that 18 436 predicted fs-proteins resulted in 'split BLASTp hits' indicative of false-positive prediction. All 206 991 fs-genes and fs-proteins were used in the analysis described below.

About 50% of fs-genes were clustered

The clustering procedure described in 'Materials and Methods' section grouped 102 731 fs-genes into 19 430 clusters. The majority of the clusters contained a small number of fs-genes: 48% contained only two fs-genes and ~75% of clusters contained less than five fs-genes. The abundance of small clusters was a result of using the stringent BLASTp threshold. Some small clusters could be related to fission events in a lineage (56).

Notably, a few clusters with up to several dozen fs-genes had similar or even identical sequences originated from closely related genomes such as genomes of 30 *E. coli* strains. Some fs-genes were detected in several copies in the same genome (e.g. genes for transposases).

Predicted programmed frameshift clusters

A cluster of fs-genes with conserved motifs located uniformly close to predicted frameshift positions was characterized as a programmed frameshift cluster. We used the GS method (see 'Materials and Methods' section) to align frameshift box sequences and identify conserved motifs. This approach, as expected, detected several known families of genes with programmed frameshifts; corresponding conserved motifs were identified.

Many known 'slippery' sequences include poly-A/T stretches [such as A_AAA_AAG (57,58) and A_A AA_AAA (59) implicated in PRF or A_n, $n > 7$ (14,40,60) and T_n, $n > 8$ (61) involved in PTR). Poly-A/T sequences

Table 3. Phylogenetic distribution of the species and genomes analyzed by GeneTack

	Taxon	Number of species	Number of genomes	Number of fs-genes	Number of fs-genes/Mb	% fs-genes in clusters
Bacteria	*Acidobacteria*	3	3	1201	60.8	9%
	Actinobacteria	72	91	24 494	58.4	29%
	Aquificae	7	7	1238	105.2	39%
	Bacteroidetes	20	22	4910	50.8	34%
	Chlamydiae	7	15	1720	94.8	84%
	Chlorobi	10	11	2118	73.5	47%
	Chloroflexi	10	14	2265	49.2	47%
	Cyanobacteria	14	38	8425	64.8	49%
	Deferribacteres	2	2	460	84.3	41%
	Deinococcus-Thermus	5	6	1179	79.8	36%
	Dictyoglomi	2	2	327	85.7	52%
	Elusimicrobia	2	2	309	111.6	16%
	Fibrobacteres	1	1	130	33.8	18%
	Firmicutes	105	199	25 890	42.4	55%
	Fusobacteria	4	4	468	43.7	19%
	Gemmatimonadetes	1	1	453	97.7	25%
	Nitrospirae	1	1	240	119.8	35%
	Planctomycetes	2	2	504	37.8	18%
	Proteobacteria	315	574	110 466	53.9	55%
	Spirochaetes	6	10	2288	73.5	77%
	Synergistetes	2	2	218	56.9	32%
	Tenericutes	5	5	436	71.7	19%
	Thermobaculum	1	2	158	50.9	25%
	Thermotogae	11	11	1350	62.1	48%
	Verrucomicrobia	4	4	690	47.1	24%
Archaea	*Crenarchaeota*	17	23	7390	146.8	60%
	Euryarchaeota	49	52	7351	59.7	27%
	Korarchaeota	1	1	174	109.4	19%
	Thaumarchaeota	1	1	139	84.5	27%
Total		680	1106	206 991		

The number of genomes for a given species depends on the number of sequenced strains.

are prone to frameshifting during translation, transcription or even replication [as DNA polymerase may produce indel errors at poly-A/T stretches (62)].

Among clusters containing at least five fs-genes, we found 146 where at least 50% of the fs-genes contained one of the seven heptamers mentioned in 'Materials and Methods' section. These clusters (with 4302 fs-genes) were divided into two groups: (i) clusters of fs-genes with known programmed frameshifts (Figure 3a) and (ii) new clusters of fs-genes predicted to use programmed frameshifts (Figure 3b).

Fs-genes with known programmed frameshifts

The Recode-2 database contains a comprehensive collection of confirmed *recoding* events, (mainly of PRF type) in prokaryotes, eukaryotes and viruses, nearly ~1500 entries (63). The recent work by Sharma et al. (2) extended the collection of known programmed frameshifts. First, Sharma et al. (2) conducted 'all against all' searches among conceptually translated protein products of disrupted coding regions annotated in GenBank. Second, the protein products were grouped into clusters of orthologs; however, the clustering did not take into account frameshift position and direction. Additionally, tBLASTn searches against the NCBI nr database were used to enrich the clusters with orthologous sequences not annotated as disrupted protein-coding regions. This approach produced 49 clusters with 8032 fs-genes.

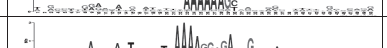
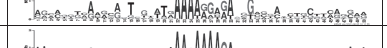
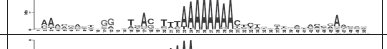
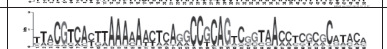
To establish correspondence between clusters identified by Sharma et al. and 146 GeneTack clusters, each of the 8032 fs-genes was used as a query in a BLASTn search against 4302 fs-genes in GeneTack clusters.

We identified 12 GeneTack clusters with fs-genes having significant sequence similarity to the fs-genes in 26 clusters of Sharma et al. We provide information on these 12 clusters in Figure 3a (11 clusters of transposase genes and a cluster of *prfB* genes).

Transposase fs-genes

In our data, genes of transposases constitute the largest group of genes with known programmed frameshifts. Interestingly, in the family of DDE_Tnp_1 transposases, we identified six clusters (three with +1 and three with −1 shift direction). Only three of them (the largest in size) matched corresponding Sharma et al. clusters. Other two clusters with −1 frameshift (of size 6 and 29) and one with +1 frameshift (6 fs-genes) could present new branches in the transposase family. In total, we identified 7 new clusters of transposase fs-genes with size ranging from 5 to 29 fs-genes. Presence of A-rich sequences in frameshift sites and existence of +1/−1 cluster pairs in a single transposase family suggest that transposase fs-genes are likely to use the PTR mechanism (see Figure 3a).

(**a**)

#	Cluster ID	Cluster name	Size	Type	Heptamer	Frameshift site Logo	Sharma et al ID
1	474411093	Release Factor 2	428 (138)	+1 (PRF)	C.TT_T.GA_C (86%)		2
2	675840861	HTH_Tnp_1 (Transposase)	1699 (106)	-1 (PTR)	AAAAAAG (49%)		11; 13; 19; 31; 33; 35
3	241541714	HTH_Tnp_1 (Transposase)	51 (12)	+1 (PTR)	AAAAAAA (51%)		12
4	667870043	HTH_Tnp_IS630 (Transposase)	495 (20)	-1 (PTR)	AAAAAAA (75%)		6; 7; 40; 62
5	858558073	HTH_Tnp_IS630 (Transposase)	185 (28)	+1 (PTR)	AAAAAAA (85%)		7
6	188472814	DDE_Tnp_1 (Transposase)	384 (37)	-1 (PTR)	AAAAAAG (28%)		5; 10; 16; 21; 38; 46; 60
7	888244788	DDE_Tnp_1 (Transposase)	108 (5)	+1 (PTR)	AAAAAAT (69%)		24
8	910763088	DDE_Tnp_1 (Transposase)	36 (1)	+1 (PTR)	AAAAAAA (100%)		10
9	696263973	DDE_Tnp_IS1 (Transposase)	230 (8)	-1 (PTR)	AAAAAAC (63%)		8
10	919140783	DDE superfamily endonuclease	43 (10)	-1 (PTR)	AAAAAAA (74%)		9
11	777059633	DDE_Tnp_ISAZ013 (Transposase)	16 (3)	-1 (PTR)	AAAAAAA (93%)		15
12	992341191	rve (Transposase)	6 (1)	+1 (PTR)	AAAAAAA (100%)		23
13	239165634	DNA polymerase III	17 (12)	-1 (PTR)	AAAAAAA (53%)		

(**b**)

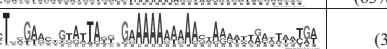
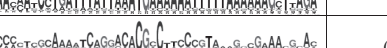
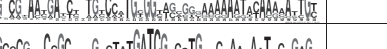
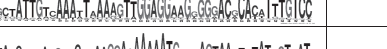
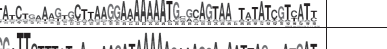
#	Cluster ID	Cluster name	Size	Type	Heptamer	Frameshift site Logo	Experimental results
1	131733585	Magnesium chelatase	23 (6)	-1 (PRF)	A_AA.A_AA.G (61%)		3/3 (63%, 40%, 10%)
2	782478235	DUF111	8 (4)	-1 (PRF)	A_AA.A_AA.A (100%)		2/2 (39%, 34%)
3	621432021	DUF772	14 (2)	-1 (PRF)	A_AA.A_AA.G (93%)		2/2 (24%, 9%)
4	447662180	Spore germination protein	19 (4)	-1 (PRF)	G_AA.A_AA.A (84%)		2/2 (13%, 4%)
5	786465964	ATP-gua_Ptrans | UVR	14 (2)	+1 (PRF+II)	C.TA_A.AA_A (50%)		1/2 (8%)
6	862991913	Phage tail assembly chaperone	41 (5)	-1 (PRF)	A_AA.A_AA.G (93%)		2/2 (7%, 6%)
7	181800409	Tetraacyldisaccharide kinase, acyltransferase	16 (4)	+1 (PRF+II)	G.AA_A.AA_A (88%)		1/2 (7%)
8	931215581	Bac_DNA_binding | Formyl_trans_N	9 (5)	+1 (PRF+II)	C.TA_A.AA_A (56%)		2/2 (6%, 3%)
9	430699271	Cyclic-nucleotide phosphodiesterase	20 (3)	-1 (PRF)	A_AA.A_AA.G (50%)		1/2 (6%)
10	720147899	phaP protein / Dehydratase (maoC family)	18 (1)	-1 (PRF)	A_AA.A_AA.G (94%)		1/2 (6%)
11	392008946	Aminotran_1_2 | Dala_lig_C | Dala_Dala_lig_N | GntR	14 (3)	+1 (PRF+II)	A.TA_A.AA_A (71%)		1/2 (6%)
12	309851863	ABC transporter	19 (3)	+1 (PRF+II)	A.AA_A.AA_A (79%)		2/2 (5%, low)
13	970108792	Preprotein translocase subunit SecA	13 (8)	+1 (II)	A.AA_A.AA_T (62%)		1/2 (3%)
14	310905921	DMRL_synthase | NusB	18 (7)	-1 (II)	C_AA.A_AA.A (56%)		1/2 (1%)
15	522343807	DNA glycosylase / Dephospho-CoA kinase	21 (2)	-1 (PRF+II)	G_AA.A_AA.A (67%)		1/2 (low)
16	984773919	Thymidylate kinase	100 (25)	+1 (II)	G.AA_A.AA_A (57%)		0/3
17	884136395	Ribosomal RNA methyltransferase	23 (5)	-1 (II)	G_AT.A_AA.A (74%)		0/2
18	645374543	Fumarylacetoacetase / Homogentisate 1,2-dioxygenase	18 (1)	+1 (II)	A.AA_A.AA_G (72%)		0/2
19	523977875	MATE efflux family protein (transporter)	11 (4)	-1 (II)	A_AA.A_AA.G (61%)		0/2
20	655521599	Epimerase | URO-D	7 (3)	-1 (II)	A_AA.A_AA.A (100%)		0/1

Figure 3. (**a**) List of GeneTack clusters corresponding to known cases of programmed frameshifting. #, row index; Cluster ID, unique identifier of a cluster (can be used for a search in GeneTack database); Cluster name, designated protein function; Size, number of fs-genes in the cluster (number of different genera is specified in parenthesis); Type, frameshift direction (possible mechanism: PTR, programmed transcriptional realignment, PRF, programmed ribosomal frameshifting, II, internal initiaion); Heptamer, overrepresented heptamer (the fraction of the cluster's fs-genes that contain the heptamer); Frameshift site Logo, logo of the frameshift site (see text for details); Sharma *et al.* ID, ID of the corresponding Sharma et al cluster(s) (2). (**b**) Summary of predicted programmed frameshifts, selected from GeneTack clusters for experimental verification. First seven column headers are the same as in Figure 3a. Experimental results (X/Y)—X programmed frameshift candidates out of selected Y candidates from a given cluster have shown detectable level of frameshifting; numbers in parentheses give frameshifting efficiency (in percentage points) for the X candidates.

prfB fs-genes

GeneTack detected frameshifts in 428 *prfB* genes encoding Release Factor 2. Expression of *prfB* uses PRF to produce full-length Release Factor 2 if its cell concentration becomes too low; regulation of *prfB* is one of the best-studied PRF instances (64). The *prfB* genes were grouped in a single cluster. The cluster could be even larger in size given that ~70% of all eubacteria are expected to use PRF to regulate expression of *prfB* (39). Still, some *prfB* genes escaped frameshift detection such as genes with frameshifts located closer than 50 nt to the start codon.

dnaX fs-genes

We have discovered a new structural type of *dnaX* genes. The GeneTack *dnaX* cluster contains 17 novel fs-genes (from 12 genera) distinctly different from seven genes (four genera) present in the Recode-2 database. None of the predicted 17 *dnaX* fs-genes with programmed frameshifts were annotated in RefSeq. These 17 *dnaX* fs-genes contain −1 frameshifts at the DNA level.

Notably, use of programmed frameshifts in the family of *dnaX* genes encoding τ and γ subunits of DNA polymerase III is well-known. The current version of the Recode-2 database (63) contains *dnaX* genes annotated as a single ORF in genomes of *Escherichia*, *Neisseria*, *Salmonella* and *Vibrio* genera. The τ subunit is the full-length product of *E. coli dnaX* gene, while the γ subunit is the shorter −1 frameshift-derived product; the N terminal regions of τ and γ subunits are identical (65–67). The *E. coli dnaX* gene was proved to use the PRF mechanism triggered by sequence A_AAA_AAG (49). Interestingly, PTR at a stretch of 9 As was shown to be used to produce γ subunit in *Thermus thermophilus* (40). Effectiveness of the *dnaX* frameshifting was shown to be 50%, which is in line with the DNA polymerase III complex stoichiometry (68).

In the *dnaX* cluster, whose 17 fs-genes span two adjacent ORFs, 12 fs-genes have poly-A stretches: nine have 10 As (see Supplementary Figure S3) and the three remaining (not shown in Supplementary Figure S3) have nine As (*Chloroherpeton thalassium*), eight As (*Chlorobium chlorochromatii*) and seven As (*Chlorobium phaeobacteroides*). The sequences with 10As can be well aligned with the poly-A motif of *T. thermophilus,* thus suggesting the same PTR mechanism for the nine species with 10As. Arguably, the *dnaX* genes with shorter poly-A motifs use PTR as well.

This type of *dnaX* genes containing two ORFs has not been described earlier.

Experimental confirmation of predicted programmed frameshifting

The remaining 134 GeneTack clusters may also contain new genes with programmed frameshifts. To experimentally verify predicted programmed frameshifts, we manually selected 40 fs-genes from 20 clusters (out of the 134 clusters) that had the most pronounced conservation around the predicted frameshift site (see Supplementary Table S2). Putative frameshift-containing sequences were cloned in vector pJ307 (see 'Materials and Methods' section). This vector has a strong promoter, pTAC, with a lac operator, the glutathione S-transferase (GST) gene lacking a terminator and fused in-frame to a maltose-binding protein (MBP) gene with a PSPXI-BamH1 cloning site between GST and MBP. The plasmid separately encodes the LacIq repressor so that expression from the pTAC promoter is inducible by addition of IPTG. The cassettes of putative frameshift-relevant sequences were inserted at the cloning site and framed to yield the fusion protein on the frameshifting in the predicted direction; the frequency of observation of the termination product synthesized without framehifting characterizes the frequency of events when ribosomes fail to change frame. The frameshift efficiency was defined as the ratio of frameshift-derived product vs total of frameshift- and non-frameshift–derived products (Supplementary Figure S4). In another experiment, we measured translational coupling (internal initiation). This test involved a His-tag encoding sequence at the 3′ end of the MBP gene with quantification by western blots with His-tag–specific antibody. The results (Supplementary Figure S5) complemented those with anti-GST western blots for frameshift identification.

We observed >10% frameshifting efficiency on testing predicted *recoding* fs-genes from four clusters: magnesium chelatase (frameshifting efficiency up to 63%), DUF111 (up to 39%), DUF772 (up to 24%) and Spore germination protein (up to 13%).

Genes for magnesium chelatase make a cluster of 23 fs-genes (with a −1 frameshift) from six different genera [in both the bacterial domain (*Pseudomonas*, *Burkholderia*, *Delftia* and *Herpetosiphon*) and the archaeal domain (*Methanocaldococcus* and *Methanococcus*)]. Cassettes made from three of the fs-genes were tested and significant levels of frameshifting, 63, 40 and 10%, were observed (Figure 4). Interestingly, the lengths of poly-A runs in the cassettes correlate with the frameshift efficiency (62): A_AAA_AAA_AAA_A (63%—11As), A_AAA_AAA_AA (40%—9As) and A_AAA_AAA (10%—7As) (Supplementary Table S2).

The newly predicted magnesium chelatase fs-genes with programmed frameshifts were annotated in RefSeq as two adjacent genes (each ~1000 nt long) with a gene for magnesium chelatase annotated upstream and some other gene annotated downstream in the gene pair. Notably, a BLASTp search against NCBI nr database reveals several magnesium chelatase proteins (from *Chloroflexus aggregans*, *Rubrobacter xylanophilus* and others) made by a fusion of the two parts, an indication that the fusion protein and the proteins whose synthesis requires a *recoding* event are likely to have similar function.

Clusters DUF111 and DUF772 were named with respect to the Pfam domains of 'unknown' type detected in these fs-proteins (with DUF standing for 'Domain of Unknown Function'). Genomic regions containing fs-genes from the DUF111 cluster are annotated in RefSeq as carrying two hypothetical genes, while in place of fs-genes from the DUF772 cluster annotation shows two separate transposases genes (IS1182 family). For fs-proteins from each cluster, similarly to the magnesium chelatase case, BLASTp searches against the

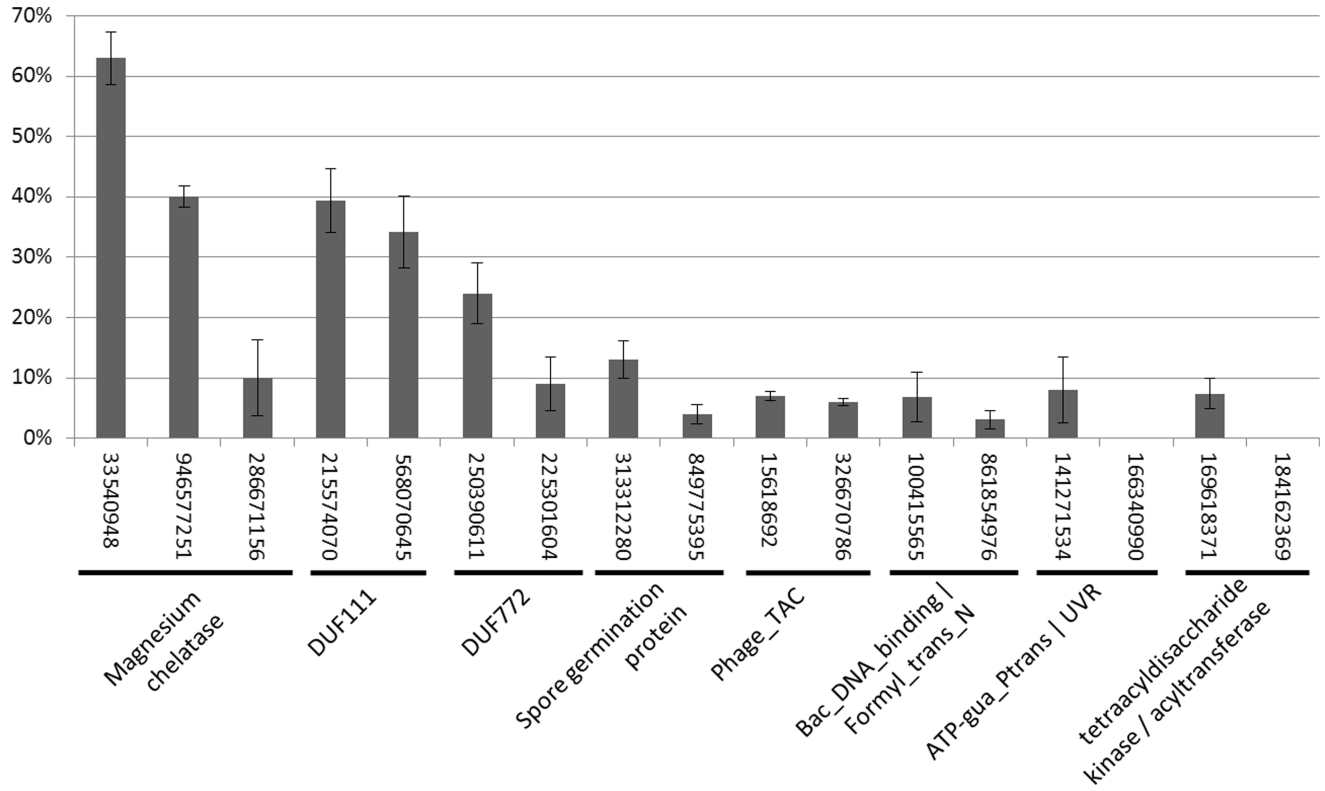


Figure 4. Experimental validation of predicted programmed frameshifting. The frameshifting efficiency in each experiment was estimated as the ratio of the product translated with the frameshift to the total amount of products translated with and without frameshift. The fs-gene ID's are listed below the graph along with the names of clusters. Note that in the last two clusters, frameshifting was observed for only one of the constructs.

NCBI nr database yield hits to fusions of two annotated proteins.

Interestingly, although the fs-genes from the Spore germination protein cluster are annotated as two separate genes encoding 'Spore germination protein', BLASTp did not produce fusion protein hits in the NCBI nr database. Still, in experiments we observed that a programmed frameshifting results in the fusion product.

We observed frameshifting efficiency of <10% in fs-genes from three clusters: phage tail assembly chaperone (7%), cyclic-nucleotide phosphodiesterase (6%) and phaP protein/dehydratase (6%). The fs-proteins from the phage tail assembly chaperone cluster (41 fs-genes) have significant similarity to a protein from *Enterobacteria phage HK97*; fs-genes encoding these fs-proteins have phage origin and are likely to use *recoding* (41,42). Still, in experiments with two fs-genes from this cluster we observed frameshifting efficiencies of 6% and 7%.

The fs-proteins from cyclic-nucleotide phosphodiesterase cluster have hits to fused proteins in the NCBI nr database suggesting similar function for products of the *recoding* fs-genes and fused genes, akin to the case of magnesium chelatase family.

Notably, a potential limitation of our experimental analysis is that frameshifting in sequences originated from different bacteria and even archaea was tested in *E. coli*. It is known that many frameshifting cassettes do not work in cross-species conditions. Thus, what we have observed, in attempt to assess the efficiency of genuine frameshiftings, is likely to give us an underestimate of the true efficiency. The fact that only parts of the genes were inserted between the reporters also contributes to producing false-negative observations. It is possible that we saw no frameshift in a particular case because the inserted sequence was too short to carry a crucial stimulatory signal. Examples of distant modulators of ribosome frameshifting are known. In *Saccharomyces cerevisiae* Antizyme mRNA modulator sequences are located at the ends of coding region (69). In Barley yellow dwarf virus a stimulatory signal was identified ~4000 nucleotides downstream of the frameshift site (70). These considerations are especially relevant for the experiments produced neither frameshifting nor initiation.

Other large clusters of fs-genes

Phase variation clusters

To identify putative phase variation clusters, we have taken the following approach. We collected a set of 38 genes with phase variation caused by the slipped strand mispairing (SSM) mechanism (53) (Supplementary Table S3). We used protein products of these genes in BLASTp searches (with E-value 10^{-10}) against the database of all fs-proteins. For the 14 queries we observed hits to 13 clusters with five or more members (Supplementary Table S4). These 13 clusters were likely to contain fs-genes with conserved phase variation. Next, we attempted to detect poly-AT stretches and short tandem

repeats characteristic for phase variation (see 'Materials and Methods' section) in the 50 nt vicinity of predicted frameshifts; the %AT and %R values were determined for each cluster (Table 2). A cluster was classified as a phase variation cluster if %AT and/or %R were higher than %S, an indicator for translational coupling (see 'Materials and Methods' section).

Translational coupling clusters

We observed that 137 clusters with five or more members contained genes for ABC transporters (4560 fs-genes), with eight clusters containing >100 members (Table 2). Earlier, it was experimentally shown that genes of ABC transporters use translational coupling, e.g. *drrAB* genes from *Streptomyces peucetius* (71), which protein products have shown similarity to fs-proteins from a GeneTack ABC transporter cluster (with 36 fs-genes). We characterized the nature of frame transitions in the ABC transporter clusters as translational coupling (Table 2).

Interestingly, the *p78* gene from the ABC transporter operon in *Mycoplasma fermentans* was characterized in (50) as a gene undergoing phase variation. However, the protein product of this *p78* gene did not have a match in our data.

Although we did not observe frameshift-derived products for several constructs used in our experiments, we did observe in such cases initiations of translation resulting in synthesis of a downstream gene products labeled with His-tag. Such observations are likely to confirm instances of translational coupling suggesting, given its conservation in fs-gene clusters, that such a co-regulation contributes to organism fitness. The clusters of fs-genes classified as translational coupling include Thymidylate kinase, Ribosomal RNA methyltransferase, Fumarylacetoacetase/Homogentisate 1,2-dioxygenase, MATE efflux family protein (transporter) and Epimerase /URO-D (Figure 3b).

Interestingly, according to experimental data, both programmed frameshifts and translation coupling may occur in fs-genes of seven clusters: 'DNA glycosylase / Dephospho-CoA kinase', 'Bac_DNA_binding/Formyl_trans_N', 'ABC transporter', 'DMRL_synthase / NusB', 'Preprotein translocase subunit SecA', 'Aminotran_1_2 / Dala_Dala_lig_C / Dala_Dala_lig N / GntR', 'ATP-gua_Ptrans / UVR' and 'Tetraacyldisaccharide kinase / acyltransferase' (Supplementary Figure S5).

Pseudogene clusters

There were 59 318 pseudogenes annotated in 1106 genomes: notably no single pseudogene was annotated in 265 genomes, while over a thousand pseudogenes were annotated in several genomes; e.g. in the parasitic bacteria *Mycobacterium leprae* (NC_011896), 1116 genes out of 2770 were annotated as pseudogenes. Notwithstanding the variability of the number of pseudogenes per genome depending on evolutionary path of a species, a low number of annotated pseudogenes in a genome could be related to far from perfect methods of pseudogene annotation.

We have found that 18 619 of the predicted fs-genes were annotated as pseudogenes and that 7186 of them belonged to clusters (3329 clusters with at least one *annotated* pseudogene). As annotation of pseudogenes may not be reliable, we excluded 411 clusters with fs-genes originated from three or more different genera (1361 fs-genes in total) assuming that real pseudogenes should be of relatively recent origin (72). Also, we excluded clusters of fs-genes with evolutionary conserved motifs in the frameshift boxes (potential clusters of fs-genes with *recoding*). We characterized the remaining 2810 clusters, each with at least one *annotated* pseudogene, as pseudogene clusters [10 290 fs-genes with 5484 fs-genes annotated as pseudogenes (Figure 5)]. Among the other 4806 fs-genes newly characterized as pseudogenes, many have appeared in genomes with no RefSeq annotated pseudogenes (Supplementary Figure S6).

Furthermore, considering fs-genes from some other clusters, we saw that predicted frameshifts should have truncated translation of a large part of an fs-gene (if not corrected by PRF or PTR). The fs-genes in these clusters did not exhibit features typical for *recoding* genes; they contained fs-genes from no more than two different genera, also, >50% of the cluster's frameshifts were validated by BLASTp (Table 1). As there were no signs of PRF or PTR mechanisms present, disfunctional truncated protein products were likely to be produced. We characterized such clusters as clusters of hypothetical pseudogenes: 1200 clusters with 3522 fs-genes (Figure 5).

A 'conserved pseudogene' may sound as a misnomer; however, with promoter and transcription process intact, its transcript could carry regulatory functions at RNA level (73) thus keeping the 'pseudogene' under selective pressure. In fact, transcription and even translation of some pseudogenes have been demonstrated experimentally (74,75).

Singletons

More than 50% of predicted fs-genes did not cluster; they formed a set of 104 260 singletons (Figure 5). Frame transitions in a singleton could be caused by sequencing error or by recent indel mutation; it may represent a false-positive fs-gene (pair of adjacent genes) as well as an orphan gene with *recoding*.

The set of singletons was divided into three groups: (i) those with frameshifts validated by BLASTp (~30% of all singletons); (ii) those with two separate BLASTp hits to proteins in other species indicating a likely false positive (~30%); (iii) those with no BLASTp hit to another protein in NCBI nr database, orphan fs-genes (~40%). We expect that singletons of category (i) represent sequencing errors rather than genuine indel mutations. Singletons of category (iii) are expected to represent gene overlaps (33%), with the remaining 67% being divided between sequencing errors and indel mutations.

We saw above that 7186 of the whole set of 18 619 annotated pseudogenes were in clusters; on the other hand, the larger fraction of this set, 11 433 annotated

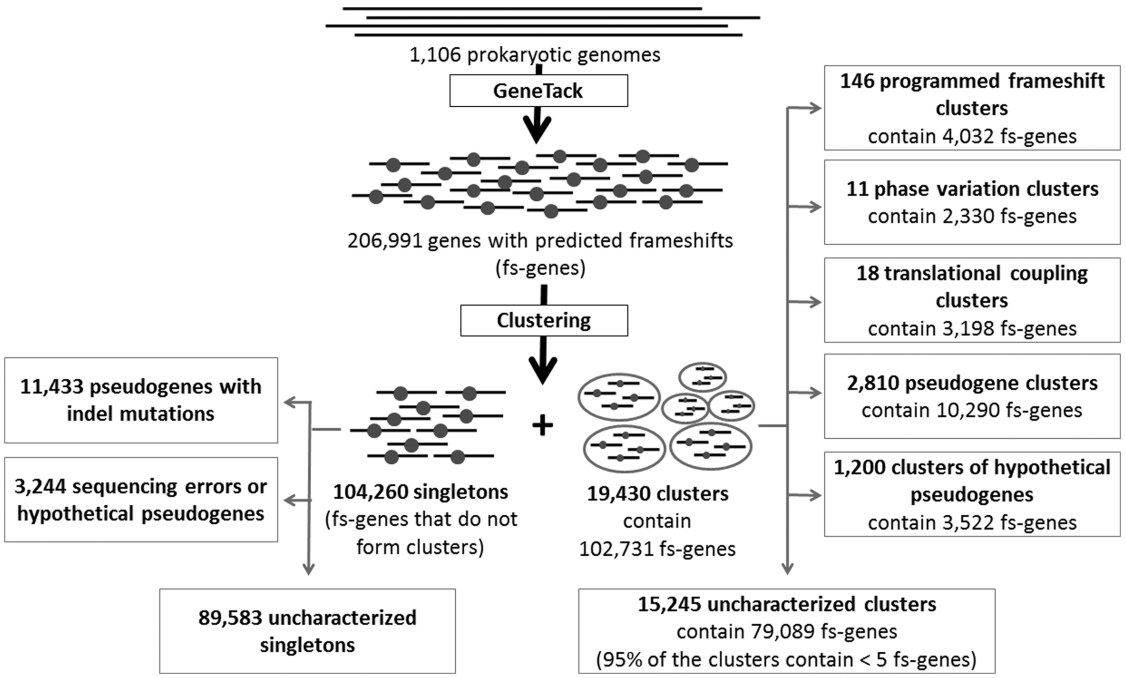


Figure 5. Classification of predicted frameshifts was done by using features specified in Table 1. One of the most important properties of a predicted fs-gene was its membership in a cluster. Singleton fs-genes (not orphan genes) are likely to be a result of indel mutation or sequencing error, while clustered fs-genes could represent programmed frameshifts, phase variation and translational coupling, as well as clusters of pseudogenes or genes with indel mutations.

pseudogenes (and predicted fs-genes) were singletons, indicating the rapid pace of pseudogenes degradation (76).

Frameshifts in 3244 singletons were confirmed by both BLASTp and Pfam; they were likely to be sequencing errors in functional genes or indel mutations in pseudogenes (see Figure 5).

Interestingly, some frameshift types are more frequent in specific locations of fs-genes (see Supplementary Materials: Distribution of relative frameshift coordinates). We observed elevation of frequency of frameshifts at the 3′-end of fs-genes (Supplementary Figure S7). One could speculate that indel mutations could truncate a gene slightly without affecting function of the protein product. Thus, a frameshift predicted close to 3′-end of a singleton would be more likely related to an indel mutation than to a sequencing error in comparison with other locations within fs-gene.

Genomic distribution of programmed frameshift sites

Sequence motifs able to trigger frameshifts, 'singular genomic elements' (3), present at specific locations close to programmed frameshift sites should be avoided at other locations. Several authors analyzed frequencies of occurrences of frameshift-prone sequences within protein coding genes. In analysis of heptamer frequencies in *S. cerevisiae* genes, Shah et al. found known frameshift-prone sequences, C.TT_A.GT_T (77,78) and C.TT_A.GG_C (79), among the least frequent heptamers (80). Our analysis of the *E. coli* genome showed that the frameshift prone motif, A_AA.A_AA.G (65,66) is underrepresented (especially in highly expressed genes); however, it is not infrequent (57). A similar pattern was observed in *H. influenzae* and *Vibrio cholerae* genomes [the sets of highly expressed genes were taken from (81)]. Interestingly, poly-AT heptamers were present even in highly expressed genes, but the poly-AT heptamer frequency *ranking* computed for highly expressed genes was always lower than the poly-AT heptamer frequency *ranking* computed for a set of genes other than highly expressed genes (see Supplementary Table S5).

DISCUSSION

Working on identification of fs-genes with programmed frameshifts we have grouped 4730 fs-genes into 146 clusters of candidate fs-genes with *recoding*. Using reporter genetic constructs based on the sequences of 20 *recoding* candidates, we confirmed that the clusters were enriched with *recoding* genes by exploring frameshifting *in vivo*. We have identified four new families of fs-genes with programmed frameshifts: fs-genes for Magnesium chelatase, Spore germination protein, DUF111 and DUF772.

While the approach to cluster characterization using multiple features (Table 1) produced a number of interesting results, the nature of frame transitions in many large clusters remained uncharacterized.

Conservation of co-location of overlapping/adjacent coding regions indicates functional relationship (82); still, the likely co-regulation of these gene pairs may use different mechanisms even in homologous genes from the

same cluster. Indeed, in some experiments we observed evidence of frameshifting along with evidence of initiation at the downstream coding region, suggesting that translational coupling and *recoding* mechanisms are not mutually exclusive but rather interchangeable. Thus, the task of unambiguous characterization of the nature of frame transition for a whole cluster may not be correctly stated.

Why does most programmed frameshifting occur in mobile elements?

A number of the largest programmed frameshift clusters were clusters of transposase fs-genes. The fact that the programmed frameshifting is so frequently used to regulate the gene expression in mobile elements is an intriguing but not entirely new observation. Contributing to selective advantage of programmed frameshifting is the fact that it is an economical gene expression regulation mechanism encoded inside the mobile element. A low, 1–3%, frameshifting efficiency moderated by stimulators around the frameshift sites results in low level of protein product. This may provide selective advantage for transposases because active fs-gene expression would result in frequent translocation of mobile elements, potentially harmful for their hosts. The low expression level would allow maintenance of a balance between proliferating/translocating and host survival (83). The poly-A slippery sequence characteristic for programmed frameshifting in transposases genes can be used in both PTR and PRF (14). Our data on transposase families where the frameshift direction (+1 or −1) is not conserved suggest that PTR is more likely to occur in this case (e.g. transposase families HTH_Tnp_1, DDE_Tnp_1, HTH_Tnp_IS630).

Interestingly, it has been shown recently that specific translation pausing during ribosomal frameshifting contributes to the preference of a transposase for acting on the IS element from which it is expressed (84). This subtle mechanism could be an additional selection force that maintains utilization of *recoding*, as the *cis*-acting transposase promotes propagation of its own mobile element and not another IS element responsive to the same transposase.

Many genes with frameshifts have incorrect annotations

Genes with frameshifts present a difficulty for standard annotation procedures. Many are incorrectly annotated either structurally or functionally. Particularly, genes with indel frameshift mutations might be annotated as two separate adjacent genes (often both genes are annotated as hypothetical genes). It is difficult to discriminate between frameshifts due to indel mutations and frameshifts related to *recoding*. Notably, several well-known genes with *recoding* were either not annotated in some genomes or annotated as pseudogenes, e.g. 17 out of 428 *prfB* genes. For these 17 *prfB* genes, a manual inspection has shown that all of them had intact programmed frameshift signals and did not have any other frameshifts or premature stop codons to justify pseudogene annotation. In the protein database, some protein sequences of Release Factor 2 were missing N-terminal ends because of wrong annotation of *prfB* genes (e.g. *Lactobacillus johnsonii NCC 533*, NC_005362). In total, out of 4302 fs-genes with predicted programmed frameshifts, 611 were annotated as pseudogenes. We are certain that at least some of them, like the *prfB* genes, were erroneously annotated. In general, due to the lack of universal methods for identification of *recoding* instances and classification of frameshifted genes, it is likely that erroneous annotations will continue to appear in databases. Nonetheless, we hope that the resource developed in the course of this research work, particularly the clusters of the fs-genes and the web-based tools of fs-gene prediction and classification will help improve annotation of frameshifted genes.

Uncharacterized clusters

In this work, we provide characterization of 38 319 fs-genes (23 642 clustered fs-genes and 14 677 singletons); this number constitutes only 19% of all predictions. Notably, 79 089 fs-genes belong to still uncharacterized clusters (Figure 5). The existence of homologous fs-genes makes it more likely that the frame transition, predicted as a frameshift, has a non-trivial biological meaning even if the transition happened between a pair of genes, not inside a single gene. These gene pairs might participate in a biological process (56) encoding functionally related proteins. Some of these gene pairs could be regulated by translational coupling, some could be formed by gene fission, etc.

The task of characterization may lead to discovery of new patterns of regulation, metabolic pathways and protein complexes. Notably, we may have missed some clusters of fs-genes with *recoding* because the 146 clusters we focused on were selected based on a limited set of the most prominent programmed frameshift motifs.

All the fs-genes predicted in this study were included in the GeneTack database (85). As of February 2013, there were 2294 genomic sequences longer than 1 Mb in the RefSeq database; 1188 new genomes have been added since the start of this project. New genomic data to be included into expanded clusters should help identify functional roles of yet uncharacterized fs-genes as well as new evolutionary conserved frame transitions.

AVAILABILITY

Additional information about fs-genes and clusters is available in the GeneTack database at http://topaz.gatech.edu/GeneTack.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–5 and Supplementary Figures 1–7.

ACKNOWLEDGEMENTS

We are grateful to Dr. Gary Loughran for his help and advice in the design of genetic constructs for testing recoding candidate sequences. We would like to thank

Alexandre Lomsadze for help in analyzing the distribution of relative frameshift coordinates and for many useful discussions. We thank Jan Mrázek for providing the sets of highly expressed genes and Karsten Suhre for addressing questions about FusionDB.

FUNDING

The work of I.A. and M.B. was supported in part by the USA National Institute of Health grant [HG000783 to M.B.]; the work of P.V.B., A.C. and J.F.A. was supported in part by the Wellcome Trust grant [094423 to P.V.B.] and by Science Foundation Ireland grant [08/IN.1/B1889 to J.F.A.]. Funding for open access charge: The Wellcome Trust grant [094423 to P.V.B.].

Conflict of interest statement. None declared.

REFERENCES

1. Antonov,I. and Borodovsky,M. (2010) GeneTack: frameshift identification in protein-coding sequences by the Viterbi algorithm. *J. Bioinform. Comput. Biol.*, **8**, 535–551.
2. Sharma,V., Firth,A.E., Antonov,I., Fayet,O., Atkins,J.F., Borodovsky,M. and Baranov,P.V. (2011) A pilot study of bacterial genes with disrupted ORFs reveals a surprising profusion of protein sequence recoding mediated by ribosomal frameshifting and transcriptional realignment. *Mol. Biol. Evol.*, **28**, 3195–3211.
3. Atkins,J.F. and Gesteland,R.F. (2010) *Recoding: Expansion of Decoding Rules Enriches Gene Expression*, 1st edn. Springer.
4. Dinman,J.D. (2012) Control of gene expression by translational recoding. *Adv. Protein Chem. Struct. Biol.*, **86**, 129–149.
5. Firth,A.E. and Brierley,I. (2012) Non-canonical translation in RNA viruses. *J. Gen. Virol.*, **93**, 1385–1409.
6. Namy,O., Rousset,J.P., Napthine,S. and Brierley,I. (2004) Reprogrammed genetic decoding in cellular gene expression. *Mol. Cell*, **13**, 157–168.
7. Baranov,P.V., Gesteland,R.F. and Atkins,J.F. (2002) Recoding: translational bifurcations in gene expression. *Gene*, **286**, 187–201.
8. Cobucci-Ponzano,B., Rossi,M. and Moracci,M. (2012) Translational recoding in archaea. *Extremophiles*, **16**, 793–803.
9. Lin,W.H. and Kussell,E. (2012) Evolutionary pressures on simple sequence repeats in prokaryotic coding regions. *Nucleic Acids Res.*, **40**, 2399–2413.
10. Kashi,Y. and King,D.G. (2006) Simple sequence repeats as advantageous mutators in evolution. *Trends Genet.*, **22**, 253–259.
11. Bekaert,M., Atkins,J.F. and Baranov,P.V. (2006) ARFA: a program for annotating bacterial release factor genes, including prediction of programmed ribosomal frameshifting. *Bioinformatics*, **22**, 2463–2465.
12. Bekaert,M., Ivanov,I.P., Atkins,J.F. and Baranov,P.V. (2008) Ornithine decarboxylase antizyme finder (OAF): fast and reliable detection of antizymes with frameshifts in mRNAs. *BMC Bioinformatics*, **9**, 178.
13. Theis,C., Reeder,J. and Giegerich,R. (2008) KnotInFrame: prediction of -1 ribosomal frameshift events. *Nucleic Acids Res.*, **36**, 6013–6020.
14. Baranov,P.V., Hammer,A.W., Zhou,J., Gesteland,R.F. and Atkins,J.F. (2005) Transcriptional slippage in bacteria: distribution in sequenced genomes and utilization in IS element gene expression. *Genome Biol.*, **6**, R25.
15. Wernegreen,J.J., Kauppinen,S.N. and Degnan,P.H. (2010) Slip into something more functional: selection maintains ancient frameshifts in homopolymeric sequences. *Mol. Biol. Evol.*, **27**, 833–839.
16. Decatur,W.A. and Fournier,M.J. (2003) RNA-guided nucleotide modification of ribosomal and other RNAs. *J. Biol. Chem.*, **278**, 695–698.
17. Byrne,E.M., Connell,G.J. and Simpson,L. (1996) Guide RNA-directed uridine insertion RNA editing *in vitro*. *EMBO J.*, **15**, 6758–6765.
18. Wulff,B.E., Sakurai,M. and Nishikura,K. (2011) Elucidating the inosinome: global approaches to adenosine-to-inosine RNA editing. *Nat. Rev. Genet.*, **12**, 81–85.
19. Kiran,A., Loughran,G., O'Mahony,J.J. and Baranov,P.V. (2011) Identification of A-to-I RNA editing: dotting the i's in the human transcriptome. *Biochemistry (Mosc)*, **76**, 915–923.
20. Maas,S. (2012) Posttranscriptional recoding by RNA editing. *Adv. Protein Chem. Struct. Biol.*, **86**, 193–224.
21. Donnelly,M.L., Luke,G., Mehrotra,A., Li,X., Hughes,L.E., Gani,D. and Ryan,M.D. (2001) Analysis of the aphthovirus 2A/2B polyprotein 'cleavage' mechanism indicates not a proteolytic reaction, but a novel translational effect: a putative ribosomal 'skip'. *J. Gen. Virol.*, **82**, 1013–1025.
22. Weiss,R.B., Dunn,D.M., Atkins,J.F. and Gesteland,R.F. (1987) Slippery runs, shifty stops, backward steps, and forward hops: −2, −1, +1, +2, +5, and +6 ribosomal frameshifting. *Cold Spring Harb. Symp. Quant. Biol.*, **52**, 687–693.
23. Atkins,J.F., Weiss,R.B. and Gesteland,R.F. (1990) Ribosome gymnastics–degree of difficulty 9.5, style 10.0. *Cell*, **62**, 413–423.
24. Atkins,J.F. and Bjork,G.R. (2009) A gripping tale of ribosomal frameshifting: extragenic suppressors of frameshift mutations spotlight P-site realignment. *Microbiol. Mol. Biol. Rev.*, **73**, 178–210.
25. Weiss,R.B., Dunn,D.M., Atkins,J.F. and Gesteland,R.F. (1990) Ribosomal frameshifting from -2 to +50 nucleotides. *Prog. Nucleic Acid Res. Mol. Biol.*, **39**, 159–183.
26. Fang,Y., Treffers,E.E., Li,Y., Tas,A., Sun,Z., van der Meer,Y., de Ru,A.H., van Veelen,P.A., Atkins,J.F., Snijder,E.J. *et al.* (2012) Efficient -2 frameshifting by mammalian ribosomes to synthesize an additional arterivirus protein. *Proc. Natl Acad. Sci. USA*, **109**, E2920–2928.
27. Baranov,P.V., Vestergaard,B., Hamelryck,T., Gesteland,R.F., Nyborg,J. and Atkins,J.F. (2006) Diverse bacterial genomes encode an operon of two genes, one of which is an unusual class-I release factor that potentially recognizes atypical mRNA signals other than normal stop codons. *Biol. Direct.*, **1**, 28.
28. Brierley,I., Gilbert,R.J. and Pennell,S. (2008) RNA pseudoknots and the regulation of protein synthesis. *Biochem. Soc. Trans.*, **36**, 684–689.
29. Giedroc,D.P. and Cornish,P.V. (2009) Frameshifting RNA pseudoknots: structure and mechanism. *Virus Res.*, **139**, 193–208.
30. Devaraj,A. and Fredrick,K. (2010) Short spacing between the Shine-Dalgarno sequence and P codon destabilizes codon-anticodon pairing in the P site to promote +1 programmed frameshifting. *Mol. Microbiol.*, **78**, 1500–1509.
31. Larsen,B., Wills,N.M., Gesteland,R.F. and Atkins,J.F. (1994) rRNA-mRNA base pairing stimulates a programmed -1 ribosomal frameshift. *J. Bacteriol.*, **176**, 6842–6851.
32. Gurvich,O.L., Nasvall,S.J., Baranov,P.V., Bjork,G.R. and Atkins,J.F. (2011) Two groups of phenylalanine biosynthetic operon leader peptides genes: a high level of apparently incidental frameshifting in decoding *Escherichia coli* pheL. *Nucleic Acids Res.*, **39**, 3079–3092.
33. Larsen,B., Peden,J., Matsufuji,S., Matsufuji,T., Brady,K., Maldonado,R., Wills,N.M., Fayet,O., Atkins,J.F. and Gesteland,R.F. (1995) Upstream stimulators for recoding. *Biochem. Cell Biol.*, **73**, 1123–1129.
34. Iseni,F., Baudin,F., Garcin,D., Marq,J.B., Ruigrok,R.W. and Kolakofsky,D. (2002) Chemical modification of nucleotide bases and mRNA editing depend on hexamer or nucleoprotein phase in Sendai virus nucleocapsids. *RNA*, **8**, 1056–1067.
35. Ferrer,I., Santpere,G. and van Leeuwen,F.W. (2008) Argyrophilic grain disease. *Brain*, **131**, 1416–1432.
36. Turnbough,C.L. Jr (2011) Regulation of gene expression by reiterative transcription. *Curr. Opin. Microbiol.*, **14**, 142–147.
37. Chamberlin,M. and Berg,P. (1962) Deoxyribo ucleic acid-directed synthesis of ribonucleic acid by an enzyme from *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **48**, 81–94.

38. Wagner,L.A., Weiss,R.B., Driscoll,R., Dunn,D.S. and Gesteland,R.F. (1990) Transcriptional slippage occurs during elongation at runs of adenine or thymine in *Escherichia coli*. *Nucleic Acids Res.*, **18**, 3529–3535.
39. Baranov,P.V., Gesteland,R.F. and Atkins,J.F. (2002) Release factor 2 frameshifting sites in different bacteria. *EMBO Rep.*, **3**, 373–377.
40. Larsen,B., Wills,N.M., Nelson,C., Atkins,J.F. and Gesteland,R.F. (2000) Nonlinearity in genetic decoding: homologous DNA replicase genes use alternatives of transcriptional slippage or translational frameshifting. *Proc. Natl Acad. Sci. USA*, **97**, 1683–1688.
41. Baranov,P.V., Fayet,O., Hendrix,R.W. and Atkins,J.F. (2006) Recoding in bacteriophages and bacterial IS elements. *Trends Genet.*, **22**, 174–181.
42. Xu,J., Hendrix,R.W. and Duda,R.L. (2004) Conserved translational frameshift in dsDNA bacteriophage tail assembly genes. *Mol. Cell*, **16**, 11–21.
43. Suhre,K. and Claverie,J.M. (2004) FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. *Nucleic Acids Res.*, **32**, D273–D276.
44. Besemer,J., Lomsadze,A. and Borodovsky,M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, **29**, 2607–2618.
45. Baranov,P.V., Gesteland,R.F. and Atkins,J.F. (2004) P-site tRNA is a crucial initiator of ribosomal frameshifting. *RNA*, **10**, 221–230.
46. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
47. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
48. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
49. Flower,A.M. and McHenry,C.S. (1990) The gamma subunit of DNA polymerase III holoenzyme of *Escherichia coli* is produced by ribosomal frameshifting. *Proc. Natl Acad. Sci. USA*, **87**, 3713–3717.
50. Theiss,P. and Wise,K.S. (1997) Localized frameshift mutation generates selective, high-frequency phase variation of a surface lipoprotein encoded by a mycoplasma ABC transporter operon. *J. Bacteriol.*, **179**, 4013–4022.
51. Srikhanta,Y.N., Fox,K.L. and Jennings,M.P. (2010) The phasevarion: phase variation of type III DNA methyltransferases controls coordinated switching in multiple genes. *Nat. Rev. Microbiol.*, **8**, 196–206.
52. van der Woude,M.W. (2006) Re-examining the role and random nature of phase variation. *FEMS Microbiol. Lett.*, **254**, 190–197.
53. van der Woude,M.W. and Baumler,A.J. (2004) Phase and antigenic variation in bacteria. *Clin. Microbiol. Rev.*, **17**, 581–611.
54. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
55. Herr,A.J., Nelson,C.C., Wills,N.M., Gesteland,R.F. and Atkins,J.F. (2001) Analysis of the roles of tRNA structure, ribosomal protein L9, and the bacteriophage T4 gene 60 bypassing signals during ribosome slippage on mRNA. *J. Mol. Biol.*, **309**, 1029–1048.
56. Szklarczyk,D., Franceschini,A., Kuhn,M., Simonovic,M., Roth,A., Minguez,P., Doerks,T., Stark,M., Muller,J., Bork,P. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.
57. Gurvich,O.L., Baranov,P.V., Zhou,J., Hammer,A.W., Gesteland,R.F. and Atkins,J.F. (2003) Sequences that direct significant levels of frameshifting are frequent in coding regions of *Escherichia coli*. *EMBO J.*, **22**, 5941–5950.
58. Wang,G., Rasko,D.A., Sherburne,R. and Taylor,D.E. (1999) Molecular genetic basis for the variable expression of Lewis Y antigen in Helicobacter pylori: analysis of the alpha (1,2) fucosyltransferase gene. *Mol. Microbiol.*, **31**, 1265–1274.
59. Mazauric,M.H., Licznar,P., Prere,M.F., Canal,I. and Fayet,O. (2008) Apical loop-internal loop RNA pseudoknots: a new type of stimulator of -1 translational frameshifting in bacteria. *J. Biol. Chem.*, **283**, 20421–20432.
60. Penno,C., Hachani,A., Biskri,L., Sansonetti,P., Allaoui,A. and Parsot,C. (2006) Transcriptional slippage controls production of type III secretion apparatus components in Shigella flexneri. *Mol. Microbiol.*, **62**, 1460–1468.
61. Penno,C., Sansonetti,P. and Parsot,C. (2005) Frameshifting by transcriptional slippage is involved in production of MxiE, the transcription activator regulated by the activity of the type III secretion apparatus in Shigella flexneri. *Mol. Microbiol.*, **56**, 204–214.
62. Kirchner,J.M., Tran,H. and Resnick,M.A. (2000) A DNA polymerase epsilon mutant that specifically causes +1 frameshift mutations within homonucleotide runs in yeast. *Genetics*, **155**, 1623–1632.
63. Bekaert,M., Firth,A.E., Zhang,Y., Gladyshev,V.N., Atkins,J.F. and Baranov,P.V. (2010) Recode-2: new design, new search tools, and many more genes. *Nucleic Acids Res.*, **38**, D69–D74.
64. Craigen,W.J. and Caskey,C.T. (1986) Expression of peptide chain release factor 2 requires high-efficiency frameshift. *Nature*, **322**, 273–275.
65. Blinkowa,A.L. and Walker,J.R. (1990) Programmed ribosomal frameshifting generates the *Escherichia coli* DNA polymerase III gamma subunit from within the tau subunit reading frame. *Nucleic Acids Res.*, **18**, 1725–1729.
66. Tsuchihashi,Z. and Kornberg,A. (1990) Translational frameshifting generates the gamma subunit of DNA polymerase III holoenzyme. *Proc. Natl Acad. Sci. USA*, **87**, 2516–2520.
67. Blinkova,A., Burkart,M.F., Owens,T.D. and Walker,J.R. (1997) Conservation of the *Escherichia coli* dnaX programmed ribosomal frameshift signal in Salmonella typhimurium. *J. Bacteriol.*, **179**, 4438–4442.
68. Larsen,B., Gesteland,R.F. and Atkins,J.F. (1997) Structural probing and mutagenic analysis of the stem-loop required for *Escherichia coli* dnaX ribosomal frameshifting: programmed efficiency of 50%. *J. Mol. Biol.*, **271**, 47–60.
69. Kurian,L., Palanimurugan,R., Godderz,D. and Dohmen,R.J. (2011) Polyamine sensing by nascent ornithine decarboxylase antizyme stimulates decoding of its mRNA. *Nature*, **477**, 490–494.
70. Paul,C.P., Barry,J.K., Dinesh-Kumar,S.P., Brault,V. and Miller,W.A. (2001) A sequence required for -1 ribosomal frameshifting located four kilobases downstream of the frameshift site. *J. Mol. Biol.*, **310**, 987–999.
71. Pradhan,P., Li,W. and Kaur,P. (2009) Translational coupling controls expression and function of the DrrAB drug efflux pump. *J. Mol. Biol.*, **385**, 831–842.
72. Lerat,E. and Ochman,H. (2004) Psi-Phi: exploring the outer limits of bacterial pseudogenes. *Genome Res.*, **14**, 2273–2278.
73. Khachane,A.N. and Harrison,P.M. (2009) Assessing the genomic evidence for conserved transcribed pseudogenes under selection. *BMC Genomics*, **10**, 435.
74. Cobucci-Ponzano,B., Guzzini,L., Benelli,D., Londei,P., Perrodou,E., Lecompte,O., Tran,D., Sun,J., Wei,J., Mathur,E.J. *et al.* (2010) Functional characterization and high-throughput proteomic analysis of interrupted genes in the archaeon Sulfolobus solfataricus. *J. Proteome Res.*, **9**, 2496–2507.
75. Feng,Y., Chien,K.Y., Chen,H.L. and Chiu,C.H. (2012) Pseudogene recoding revealed from proteomic analysis of salmonella serovars. *J. Proteome Res.*, **11**, 1715–1719.
76. Kuo,C.H. and Ochman,H. (2010) The extinction dynamics of bacterial pseudogenes. *PLoS Genet.*, **6**
77. Morris,D.K. and Lundblad,V. (1997) Programmed translational frameshifting in a gene required for yeast telomere replication. *Curr. Biol.*, **7**, 969–976.
78. Taliaferro,D. and Farabaugh,P.J. (2007) An mRNA sequence derived from the yeast EST3 gene stimulates programmed +1 translational frameshifting. *RNA*, **13**, 606–613.
79. Asakura,T., Sasaki,T., Nagano,F., Satoh,A., Obaishi,H., Nishioka,H., Imamura,H., Hotta,K., Tanaka,K., Nakanishi,H. *et al.* (1998) Isolation and characterization of a novel actin

filament-binding protein from Saccharomyces cerevisiae. *Oncogene*, **16**, 121–130.
80. Shah,A.A., Giddings,M.C., Parvaz,J.B., Gesteland,R.F., Atkins,J.F. and Ivanov,I.P. (2002) Computational identification of putative programmed translational frameshift sites. *Bioinformatics*, **18**, 1046–1053.
81. Karlin,S., Mrazek,J., Campbell,A. and Kaiser,D. (2001) Characterizations of highly expressed genes of four fast-growing bacteria. *J. Bacteriol.*, **183**, 5025–5040.
82. Wolf,Y.I., Rogozin,I.B., Kondrashov,A.S. and Koonin,E.V. (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.*, **11**, 356–372.
83. Nagy,Z. and Chandler,M. (2004) Regulation of transposition in bacteria. *Res. Microbiol.*, **155**, 387–398.
84. Duval-Valentin,G. and Chandler,M. (2011) Cotranslational control of DNA transposition: a window of opportunity. *Mol. Cell*, **44**, 989–996.
85. Antonov,I., Baranov,P. and Borodovsky,M. (2013) GeneTack database: genes with frameshifts in prokaryotic genomes and eukaryotic mRNA sequences. *Nucleic Acids Res.*, **41**, D152–D156.