High-throughput interrogation of programmed ribosomal frameshifting in human cells

Martin Mikl[123*], Yitzhak Pilpel[3] and Eran Segal[12*]

[1]Department of Computer Science and Applied Mathematics, [2]Department of Molecular Cell Biology and [3]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 7610001, Israel.

*Correspondence: eran.segal@weizmann.ac.il, martin.mikl@weizmann.ac.il

Summary

Programmed ribosomal frameshifting is the controlled slippage of the translating ribosome to an alternative frame. This tightly regulated process is widely employed by human viruses such as HIV and SARS coronavirus and is critical for their life cycle and virulence. It is also utilized from yeast to human to implement a feedback control mechanism to regulate polyamine levels. Here, we developed a high-throughput, fluorescence-based approach to assess the frameshifting potential of a sequence. We designed and tested >12,000 sequences based on 15 viral and human frameshifting events, allowing us to elucidate the rules governing ribosomal frameshifting in a systematic way and to discover novel regulatory inputs based on amino acid properties and tRNA availability. We assessed the natural variation in HIV gag-pol frameshifting rates by testing >500 clinical isolates and identified subtype-specific differences as well as associations between viral load in patients and the optimality of gag-pol frameshifting rates. We further devised computational models that accurately predict frameshifting potential (up to auROC=0.93) and frameshifting rates (up to Pearson r=0.81) of novel variants, including subtle differences between HIV clinical isolates (r=0.60). Taken together, this systematic approach can contribute to the development of antiviral agents acting on programmed ribosomal frameshifting.

Introduction

Programmed ribosomal frameshifting (PRF), i.e. controlled slippage of the ribosome, is a mechanism by which two proteins with alternative C termini can be generated from the same mRNA. It allows for an expansion of the proteome but also constitutes an additional regulatory layer to fine-tune gene expression (Advani and Dinman, 2016; Dinman, 2012; Ketteler, 2012). This mechanism is widespread and indispensable in viruses, which often utilize controlled slippage of the ribosome to an alternative frame to regulate the production of key enzymes, such as in the case of the gag-pol frameshift in HIV and other retroviruses. Utilizing alternative frames increases the amount of genetic information that can be encoded in a given sequence and constitutes another level of gene regulation, which might explain why especially RNA viruses – with their compact genomes and lack of regulation of their genes on the transcriptional level - utilize PRF for crucial regulatory switches (Firth and Brierley, 2012). The importance of maintaining the stoichiometry between structural proteins encoded by the *gag* gene and enzymes encoded by the *pol* gene for viral replicative success make the gag-pol frameshifting event a promising antiviral drug target (Brakier-Gingras et al., 2012; Hung et al., 1998).

Cases of functionally important programmed frameshifting have also been discovered in humans (Belew et al., 2014; Clark et al., 2007; Ivanov et al., 1998; Matsufuji et al., 1995; Rom and Kahana, 1994; Tosaka et al., 2000; Wills et al., 2006). Discovering PRF events in the human genome has been hampered by the limited amenability of PRF to proteome-wide methods due to the generally low abundance of the frameshifted protein relative to the canonical protein or inherent instability of the frameshifting product. Most human cases known to date were found serendipitously or through homologous genes. A striking example of regulatory conservation is the case of ornithine decarboxylase antizyme (OAZ), which is produced through polyamine-stimulated +1 frameshifting and inhibits polyamine production (Matsufuji et al., 1995; Rom and Kahana, 1994). This negative feedback loop is used by virtually all organisms from yeast to humans to control polyamine levels (Ivanov et al., 2000), attesting to the evolutionary success of PRF as a regulatory mechanism.

Fameshifting is generally believed to happen at defined positions consisting of a slippery sequence and a downstream roadblock, most commonly a stable secondary RNA structure like a pseudoknot or an extensive stem-loop structure (Caliskan et al., 2015; Dinman, 2012). Most presently known -1 slippery sites follow the pattern X XXY YYZ, with the shift happening from codons XXY and YYZ to XXX and YYY. Some known slippery sites in -1 PRF show divergence from this pattern, e.g. in the NSA2 gene of west nile virus with the slippery site sequence UCCUUUU (Firth and Atkins, 2009; Melian et al., 2010). At +1 frameshifting sites like OAZ ribosomal translocation happens at a very distinct motif (e.g. UCCUGA). Many case studies have contributed to an understanding of the molecular events happening during frameshifting (e.g. Belew et al., 2014; Caliskan et al., 2015; Kurian et al., 2011; Ritchie et al., 2017; Tholstrup et al., 2012), but the general, overarching regulatory principles that determine if and to what extent PRF occurs remain largely unknown.

62 Here, we developed a massively parallel reporter assay that allows for high-throughput quantification
63 of ribosomal frameshifting in human cells. We designed and tested 13,390 oligonucleotides containing
64 rationally designed variants of known frameshifting signals. We systematically deciphered
65 determinants of PRF efficiency across frameshifting events and assayed natural variation in HIV gag-
66 pol frameshifting, providing the first systematic large-scale investigation of ribosomal frameshifting.

67

68 **Results**

69 *A massively parallel reporter assay for programmed ribosomal frameshifting*

70 To assay PRF in a comprehensive manner we designed a synthetic oligonucleotide library containing
71 12809 variants with systematic sequence manipulations of previously reported PRF sites
72 (Supplementary Data 1, "main set") and 581 sequences of gag-pol frameshifting sites in HIV clinical
73 isolates (collated from http://www.hiv.lanl.gov/, see Methods for selection criteria of HIV variants) (Fig
74 1A).

75 The oligonucleotides comprising library-specific common primers, a unique barcode and a 162 nt long
76 variable region containing a potential frameshifting site (Fig 1A, bottom) were synthesized on an
77 Agilent microarray, amplified and cloned in between *mcherry* and *gfp* coding sequences such that the
78 *gfp* coding frame was shifted by +1 or -1 relative to the original, mCherry-encoding frame (Fig 1B). If
79 the corresponding frameshift occurs, GFP is made into protein, and GFP fluorescence intensity thus
80 serves as a measure for frameshifting efficiency. We introduced this construct in the AAVS1 locus in
81 the human K562 cell line using zinc finger nucleases, such that every cell has one frameshifting
82 reporter construct from the library and all the variants have the same genomic environment (Methods).
83 For both, the -1 and +1 reporter libraries, we selected a narrow mCherry-positive population to
84 minimize effects coming from the influence of the variable region on overall expression levels (Fig
85 S1AB, Methods). We sorted this population corresponding to a single integration of the reporter
86 transgene using flow cytometry into 16 bins according to their GFP fluorescence intensity (Fig S1CD)
87 and sequenced genomic DNA from all the bins to determine the distribution of each variant across
88 bins. We previously demonstrated that similar approaches are highly accurate and reproducible (Mikl
89 et al., 2019; Vainberg Slutskin et al., 2018; Weingarten-Gabbay et al., 2016, 2019), and the consistent
90 bin profiles for many barcode control groups with identical variable region (Fig S1E, Methods)
91 corroborate the low technical noise we are able to achieve.

92 The distribution of mean GFP expression of all library variants showed a clear peak corresponding to
93 background green fluorescence (Fig S2AB). We set the lowest mean GFP fluorescence we observed
94 to 0% and the highest – coming from variants with GFP in frame with mCherry – to 100%. Accordingly,
95 we assigned a percentage to every variant that passed filtering for read number, bin profile and

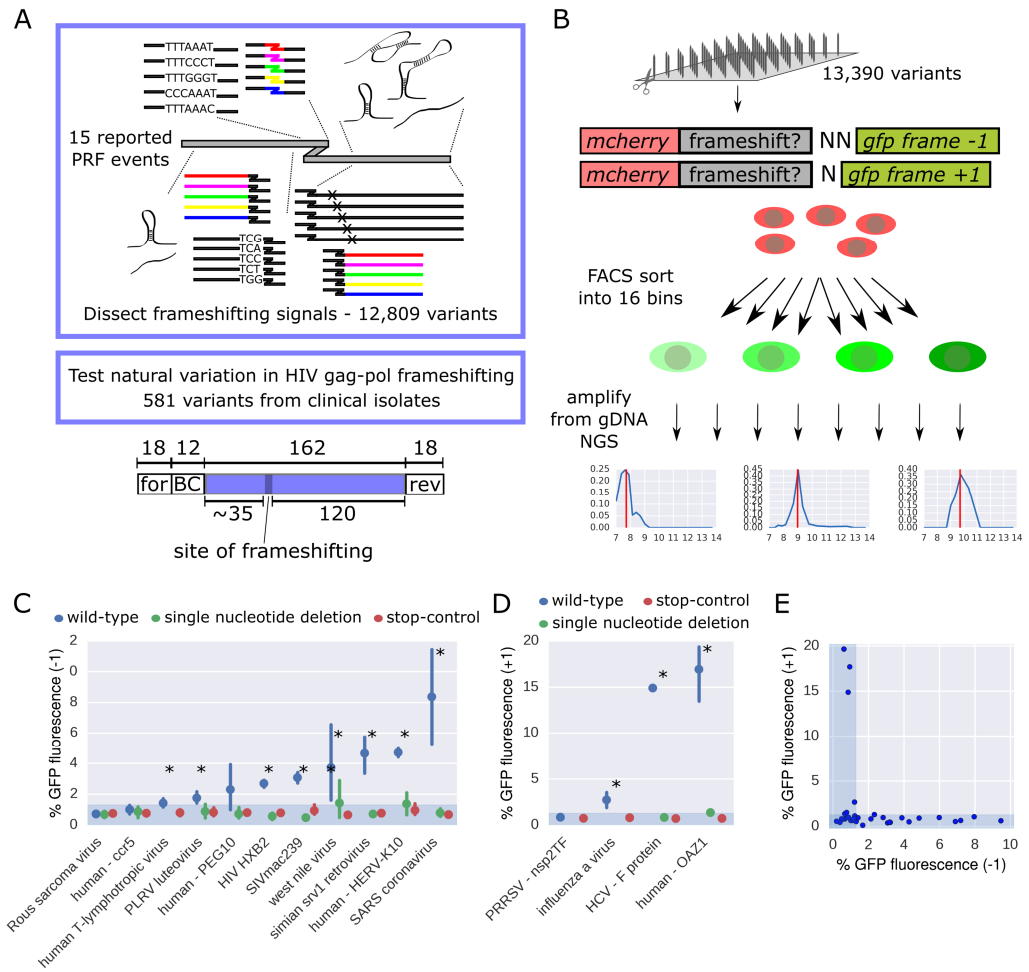

Figure 1. A massively parallel reporter assay for programmed ribosomal frameshifting.

A. Schematics of the library design; for: forward primer, rev: reverse primer, BC: barcode. B. Outline of the experimental pipeline; the plots in the bottom constitute representative bin profiles for variants with no (left), low (middle) or intermediate (right) GFP fluorescence. CD. Mean and 95% CI of -1 (C) and +1 (D) PRF signal (% GFP fluorescence) for barcode control groups corresponding to the indicated wild-type PRF sequences (blue), carrying single nucleotide deletions (green) or having a stop codon inserted upstream of the frameshift sequence (red); n between 1 and 32; the shaded area denotes the range of background fluorescence; asterisks denote significant differences between the wild-type and the other groups (single nucleotide deletion and stop-control) combined (Mann-Whitney-U test). E. Comparison of -1 and +1 PRF reporter readout (% GFP fluorescence) for previously reported PRF sites (cf. Fig 1CD, Fig S3AC).

expression levels (by gating for a narrow range of mCherry fluorescence to minimize effects coming from the influence of the variable region on overall expression levels, Methods). This percentage value does not denote the precise rate of frameshifting events, but gives us a meaningful measure of frameshifting efficiencies across PRF events. We thereby obtained measurements for 8972 (67%) and 5922 (44%) sequences cloned into the -1 PRF and +1 PRF reporter, respectively.

102 To test the possibility that the signal in our assay comes from internal translation initiation downstream
103 of mCherry (which might lead to GFP expression independently of ribosomal frameshifting in between
104 the *mCherry* and *gfp* coding region) we cloned our library in a modified reporter construct containing
105 a stop codon after the mCherry coding region (Fig S2C top). GFP fluorescence coming from a
106 frameshifting event would in this case be lost. We subjected this library to the same experimental
107 pipeline and obtained green fluorescence measurements for 10561 variants (79% of the library, 82%
108 and 83% of variants for which we obtained -1 and +1 PRF readouts, resp.; Fig S2C). Green
109 fluorescence was almost exclusively at background levels, showing that the overwhelming majority of
110 tested sequences cannot drive GFP expression on their own.

111 Another potential source of GFP expression is the removal of cryptic introns in the variable region that
112 would lead to *gfp* being in frame with *mCherry* in the resulting mature mRNA. To test if this is a source
113 of false positives in our assay, we performed RNA sequencing on the whole library, covering the entire
114 variable region and its surroundings (Fig S2D). We mapped the full length reads to the library variants
115 to detect splicing events and obtained an RNA readout (at least 100 reads mapped) for 10666 variants
116 (80% of the library, 94% of the variants with readout in the -1 frame and 96% of the variants with
117 readout in the +1 frame). Even without any filtering, only 2.95% of the variants showed gaps in the
118 mapping of at least 20 nt which could represent events of cryptic splicing (Fig S2EF; 0.95% introducing
119 a shift to frame +1 and 0.64% to frame -1). Missing stretches can also come from synthesis or cloning
120 errors (and are therefore present and filtered out on the DNA level. Considering only those cases that
121 had a potential (degenerate, cryptic) donor or acceptor splice site in the area of the gap further
122 reduced the fraction of variants for which we cannot rule out a relevant splicing event to 36 (0.55%;
123 Fig S2EF, red; 0.22% introducing a shift to frame +1 and 0.14% to frame -1). Importantly, even for
124 these cases there was no correlation between the fraction of potentially spliced reads and GFP
125 expression (Pearson r=0.01, p=0.21 for -1 PRF and r=0.02, p=0.09 for +1 PRF), indicating that cryptic
126 splicing events do not constitute a common source for false positive signal in our assay.

127 During oligonucleotide synthesis errors can occur that can introduce a point mutation or an
128 insertion/deletion. For assaying ribosomal frameshifting, especially the latter type would be
129 detrimental, and therefore we perform full-length DNA sequencing in our assay and consider only
130 reads that exhibit perfect alignment along the length of the variable region covered by sequencing
131 reads (approximately 20 nt in the center of the variable region cannot be covered due to the read
132 length (150 nt from either side)). These synthesis errors, however, also further enrich our synthetic
133 library, as they add additional variants to the collection of sequences tested. Especially single
134 nucleotide deletions or insertions constitute a great internal control, as they differ only in one position
135 from the designed variant. Cases where we obtained enough reads (after filtering, Methods) for the
136 intended variant (Fig S2G top, “wild-type”), a single nucleotide deletion abolishing the signal coming
137 from -1 PRF (Fig S2G middle) and a single nucleotide insertion leading to *gfp* being in frame (Fig S2G

138 bottom) constitute a proof of principle. They also provide additional evidence that the signal that we
139 measure stems from a (reading frame-sensitive) ribosomal frameshifting event. Looking at the
140 distribution of all single nucleotide deletion and insertion variants from the library reveals the expected
141 peaks at minimal and maximal GFP expression (Fig Fig S2HI), with a small group of variants being
142 misclassified due to the fact that our reads cover only approximately 90% of the variable region.

143 Based on the distribution of all library variants with a stop codon cassette integrated immediately after
144 the *mcherry* coding sequence we assigned a noise threshold (95th percentile of the stop-vector library
145 distribution, corresponding to 1.3% of the maximal GFP fluorescence; Fig S2C) imposed by
146 autofluorescence of the cells. The contribution of autofluorescence coming from cellular components
147 and metabolites such as NADH and flavins is negligible when methods for signal amplification are
148 used (e.g. in Luciferase assays commonly used for the quantification of PRF). It does, however,
149 constitute a limitation for detection of GFP-fusion proteins made from a single gene copy and present
150 at low copy number in the cell owing to the fact that the PRF event responsible for its expression
151 happens in only a few percent of translation events. This entails a higher detection limit of our assay
152 compared to for example Luciferase assays, which is a necessary compromise to enable high-
153 throughput testing.

154 Based on the distribution of GFP intensities of variants with a mutation leading to GFP being in frame
155 with mCherry (Fig S2I) we also excluded variants exhibiting a dominant peak with mean GFP
156 fluorescence above 2^{12} (25% of maximal GFP fluorescence). This value corresponds to the 5th
157 percentile of the distribution of variants with a single nucleotide insertion (above threshold). We
158 thereby aim to rule out biases stemming from DNA frameshifts occurring as synthesis or cloning errors
159 that our mapping strategy might have missed (Methods).

160 We tested previously reported frameshifting sites (with or without experimental validation,
161 Supplementary Data 1, based on Moon et al., 2007 and our own survey of the literature) with multiple
162 different barcodes in our assay and could reproducibly detect GFP fluorescence in the expected frame
163 for many of the previously reported PRF events (Fig 1CD, Fig S3A-D). For example, the previously
164 reported PRF sites in HIV-1 (Jacks et al., 1988a), SARS coronavirus (Thiel et al., 2003), HERV-K10
165 (Ono et al., 1986; Wang et al., 2002), simian retrovirus (ten Dam et al., 1994), SIVmac239
166 (Marcheschi et al., 2007; Pollom et al., 2013), human T cell lymphotropic virus (Nam et al., 1993) and
167 OAZ (Matsufuji et al., 1995; Rom and Kahana, 1994) yielded a fluorescent signal between 1.4% and
168 18% of maximal GFP fluorescence detected in our assay. These values do not necessarily denote
169 the percentage of translation events, in which frameshifting happens, in absolute terms, and are not
170 directly comparable with percentages reported in other studies.

171 PRF events tested in our assay typically gave a signal over threshold only in either the -1 or +1 PRF
172 reporter, but not in both (Fig 1E). In cases where the native sequence contained stop codons in any

173 frame downstream of the frameshifting site, we also tested variants with all stop codons mutated
174 (Methods). These typically either did not affect or led to a substantially reduced frameshifting signal
175 (Fig S4A). Moving the slippery site by one to three nucleotides typically abolished the signal, e.g. in
176 the case of SARS (Fig S4B, Methods).

177 Some previously reported PRF events failed to yield a signal above threshold in our assay. Several
178 reasons like sensitivity to the larger context of the reporter construct or a general effect of the tested
179 sequence on expression could be possible explanations. In the case of the PRF site in Rous sarcoma
180 virus (Jacks and Varmus, 1985; Fig S4C), a construct containing 124 nt of sequence downstream to
181 the position of frameshifting was able to induce frameshifting, but shorter versions were not (Jacks et
182 al., 1988b). A sequence motif (termed PK4) at the end of this downstream region was shown to be
183 important for maintaining wild-type frameshifting ability, potentially by creating a pseudoknot-like
184 structure through basepairing with a complementary sequence in the loop of the main stem-loop
185 structure (Marczinke et al., 1998; Nikolic et al., 2012). Our construct (including 120 nt of downstream
186 sequence) does contain the PK4 motif, but it is located at the very 3' end of the native sequence
187 included in the reporter construct. The sequence context of PK4 was shown to be important for proper
188 folding (Marczinke et al., 1998), and therefore the specific context created by our reporter construct
189 might explain the lack of signal.

190 We also failed to detect a frameshifting signal for CCR5 (Belew et al., 2014; Fig S4C). In addition to
191 potential sensitivity to the larger context of the reporter construct and a specific miRNA that is required
192 for maximal frameshifting efficiency (Belew et al., 2014), the lack of signal might be due to a potential
193 effect of the PRF site on overall expression levels on the RNA or protein level, as has been suggested
194 for cellular frameshifting sites in general (Advani and Dinman, 2016). We select for a narrow range of
195 mCherry levels and therefore expression levels (Methods, Fig S1AB), which leads to filtering out
196 variants that strongly affect overall expression levels. In the case of CCR5 only 29% of variants with
197 the wild-type sequence passed filtering (Fig S4D; as opposed to 45-97% for other PRF events), which
198 could hint at an effect of the included sequence on expression at the RNA or protein level. Estimating
199 steady state RNA levels did not show systematic differences between PRF events yielding a signal
200 above threshold in our assay and those that failed to do so (Fig S4E), but indicated that specifically
201 RNA from variants containing the PRRSV PRF site was present at reduced levels.

202 Some of the frameshifting signals tested showed variability between variants with identical variants,
203 but different barcodes (Fig S4C, right). In the case of west nile virus, the majority of sequences tested
204 resulted in a fluorescent signal above threshold, leading to it being classified as "frameshifting" in our
205 assay ($p=0.027$ for the difference between wild-type and stop control; Fig 1C); in the case of PEG10
206 (Clark et al., 2007; Manktelow et al., 2005) the majority of sequences failed to yield a signal (after
207 filtering). This points at an instability of the signal in this context, and therefore our assay might not be

208 suitable for investigating these specific events. PEG10 is thought to have a relatively extensive
209 downstream region necessary for frameshifting (~80 nt according to estimates from in vitro testing of
210 truncated regions (Manktelow et al., 2005)). Our library includes 120 nt of downstream sequence, but
211 the more extensive the downstream sequence, the higher the sensibility to the context and
212 experimental conditions might be. This is in line with our observation for SARS and HERV-K10, for
213 example (see below). In addition, the required context might be larger in the *in vivo* situation compared
214 to *in vitro*.

215 This instability of the signal might be affected by small sequence changes: For some library
216 sequences measurements of a large number of variants with sequence alterations introduced during
217 synthesis and cloning are available. In the case of PEG10, all the synthesis variants containing a
218 specific mutation (C-to-G at position 7 after the slippery site) show a -1 PRF signal above background
219 (Fig S4F). Likewise, a fraction of the designed sequence alterations introduced in the PEG10 context
220 triggered frameshifting (Fig S4G), especially mutations in the upstream region (Fig S4G left). This
221 shows that minimal sequence changes can have far-reaching (and consistent) consequences for the
222 potential of a sequence to yield a frameshifting signal in the context of our high-throughput assay and
223 probably in any experimental setup.

224

225 *Slippery sites are optimized for the respective frameshifting event*

226 While previous investigations of frameshifting focused on individual examples, we aimed to identify
227 commonalities and differences between frameshifting sites. To this end we introduced systematic
228 sequence alterations (such as point mutations, synonymous substitutions, altering and mimicking
229 endogenous secondary structure, etc.; Fig 1A; Methods) in all contexts and measured the effect on
230 frameshifting in large scale using our FACS based assay. Individual PRF events exhibited
231 characteristic sensitivities to sequence alterations (Fig 2A). The HIV-1 gag-pol PRF site (Jacks et al.,
232 1988a) was particularly sensitive to changes in or immediately around the slippery site (any sequence
233 change in this region led to a median PRF signal of 25% of the wild-type), while changes in the
234 upstream or downstream region led to PRF signal of – on average – 97% (upstream) and 70%
235 (downstream; Fig 2A, Fig S5A) of the corresponding wild-type. A similar pattern could be observed
236 for SIVmac239. The HERV-K10 (Ono et al., 1986; Wang et al., 2002), SARS (Thiel et al., 2003) and
237 SRV1 (ten Dam et al., 1994) PRF sites were highly sensitive to changes in both the slippery site and
238 the downstream region and in the case of HERV-K10 and SRV1 also showed a significant (p=0.0004
239 and 0.0023, resp.; Wilcoxon signed-rank test) reduction following changes in the upstream region (Fig
240 2A, Fig S5B). It is noteworthy that the clustering of effects (although done on percent of wild-type PRF

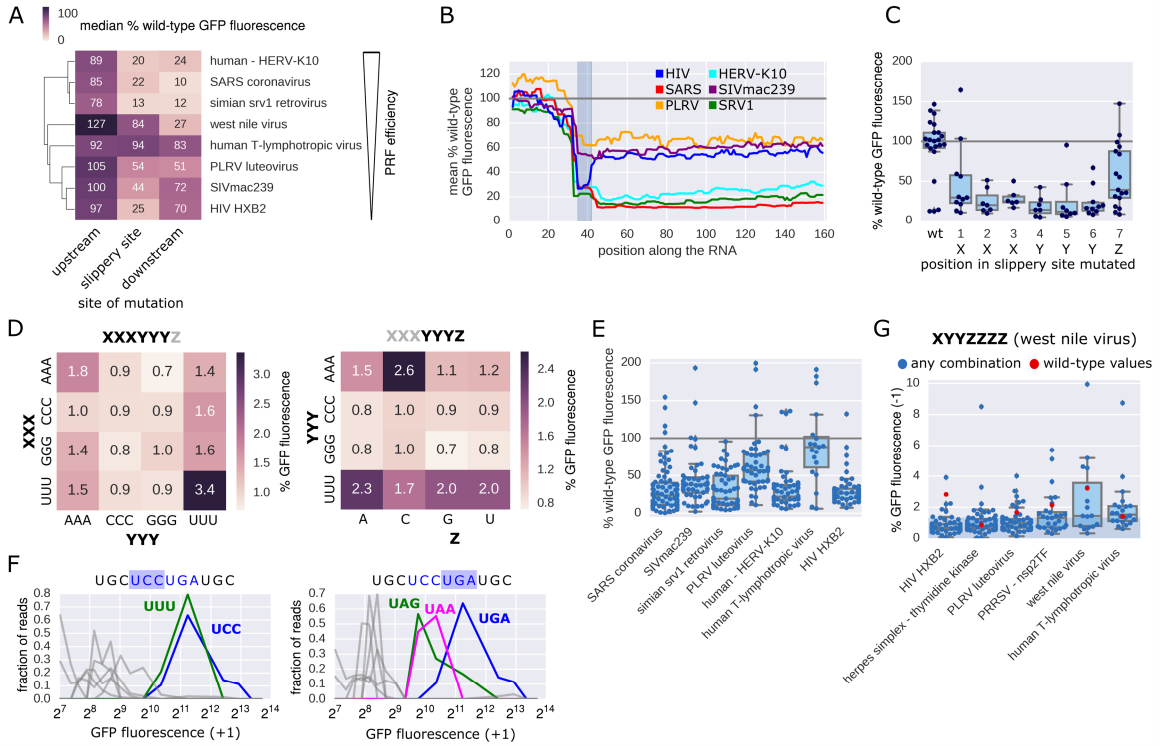

Figure 2. Slippery sites are optimized for the respective frameshifting event.

A. Clustered heat map of median percent wild-type frameshifting of variants in which the indicated region (upstream, slippery site (including the preceding codon) and downstream) differs from the native one (n=37-325). B. The median percent wild-type PRF signal of variants in which the indicated position is changed is plotted for the entire length of the variable region; gray box: slippery site. C. Percent wild-type -1 PRF signal for variants with a point mutation only in the indicated position of the slippery site (n=6-22). D. Heatmap showing median percent GFP fluorescence conferred by replacing the native slippery site in -1 PRF sites tested positive in the assay (Fig 1C) with the indicated combination of slippery site elements. E. Distribution of percent wild-type GFP fluorescence of variants in which the slippery site was replaced by any possible combination of bases following the pattern XXXYYYZ (n=19-80). F. Distribution of normalized reads across GFP (+1 frame) expression bins for variants of the OAZ1 frameshifting site, in which either the first (right) or the second (left) codon is replaced; blue lines: wild-type; green and magenta lines: bin profile for the variant in which the native codon was replaced by the indicated one; grey lines: bin profiles for other codons at this position. G. Percent GFP fluorescence for variants in which the slippery site was replace with all possible combinations of the slippery site pattern found in west nile virus (XYYZZZZ) (n=16-56).

signal) resembles the wild-type PRF efficiencies (Fig 1C), with the most efficient sites (HERV-K10, SARS and SRV1) showing the highest and broadest sensitivity to sequence alterations.

Comparing per position the median PRF signal for variants in which this position was changed reveals a sensitivity profile around the PRF site (Fig 2B, Fig S5C), recapitulating these context specific characteristics. Across all -1 PRF events tested positive in our assay, not only mutations at the slippery site (marked in grey) and downstream of it, but also mutations immediately upstream of the slippery site negatively affect PRF efficiency. Other -1 PRF events reported in the literature, but not yielding a signal above background in our assay, tend to show substantially increased signal upon

249 mutations in the areas upstream (PEG10) and downstream (PEG10, CCR5, Rous sarcoma virus) of
250 the slippery site (Fig S5D). Like the specific examples of frameshifting PEG10 variants described
251 above (Fig S4EF), this points at a strong sensitivity of these PRF events to the context in which they
252 are tested.

253 Most single point mutations within the slippery sequence abrogated frameshifting and resulted in
254 green fluorescence at background levels, both for -1 and +1 frameshifting sites (Fig 2C, Fig S5E),
255 demonstrating that we indeed measure frameshifting at the expected site. Replacing the last position
256 in the canonical slippery site X XXY YYZ had – on average – a less severe effect on PRF efficiency,
257 due to the fact that it is the wobble position in the original frame and therefore has larger flexibility (Fig
258 2C). In high efficiency PRF sites also this position showed strong and PRF site-specific preferences:
259 In the case of SRV1 (GGGAAAC) replacing the Z position with other options than the endogenous C
260 led to a drastic decrease (U, A) or complete loss (G) of -1 PRF signal (Fig S4E), while in the case of
261 SARS (UUUAAAU) all other nucleotides apart from the endogenous U at the Z position almost
262 completely abolished the signal. To determine whether there is a universally optimal slippery site, we
263 replaced the slippery sequence in our set of -1 PRF events with all possible variations of the pattern
264 XXXYYYZ (Fig 2D, Fig S4F). This revealed preferences for specific combinations common in known
265 PRF sites (Fig 2D), like UUUUUUZ (HIV, SIVmac239) and XXXAAAC (HERV-K10, HTLV, PEG10,
266 SARS), but no universally optimal combination of bases that would lead to maximal frameshifting
267 efficiency and consequently GFP fluorescence across contexts (Fig S5F). Notably, PRF sites showed
268 a high degree of optimization when it comes to the identity of the bases, with any substitution (although
269 preserving the canonical XXXYYYZ pattern) typically leading to a more or less severe drop in PRF
270 signal (Fig 2E, Fig S5EF).

271 Replacing the first or second codon of the OAZ slippery site (UCC UGA) revealed a very limited
272 tolerance to sequence alterations, consistent with earlier results (Matsufuji et al., 1995). Replacing
273 the second codon (UGA), the only variants leading to a +1 PRF signal over background where the
274 other two stop codons (UAA and UAG; Fig 2F, right), reaching approximately 50% of wild-type
275 frameshifting rates, confirming that a translation termination signal is required for frameshifting (Rom
276 and Kahana, 1994). At the first codon position, UUU was the only codon leading to +1 PRF signal
277 similar to wild-type (Fig 2F, left), maybe due to the creation of a stretch of four Us and thereby
278 facilitating re-pairing of the tRNA in the +1 frame. Indeed, UUU is the second most widely used codon
279 found in that position in the OAZ frameshifting site in other species, e.g. certain funghi and nematodes
280 (Ivanov and Atkins, 2007).

281 Some reported -1 PRF sites deviate from the canonical slippery site pattern (XXXYYYZ). To test the
282 ability of the alternative patterns to induce frameshifting across sequence contexts we replaced the
283 slippery site in a set of PRF events with all combinations of nucleotides following the pattern found in

284 west nile virus (XYYZZZZ, Fig 2G; Melian et al., 2014; Moomau et al., 2016). In general, the ability of
285 alternative slippery site patterns to cause frameshifting was limited to their native sequence contexts.
286 In this context, however, several different nucleotide combinations following the respective pattern
287 were tolerated. The only exception to this rule was human T-lymphotropic virus (Fig 2G, endogenous
288 slippery site UUUAAAC), which gave a robust -1 PRF signal corresponding to wild-type levels across
289 many variants following the (west nile virus) pattern XYYZZZZ. In summary, our data suggest that in
290 most cases a slippery site not following the canonical pattern requires a specific sequence context in
291 order to trigger frameshifting, in accordance with earlier results comparing west nile virus strains with
292 canonical or non-canonical slippery sites (Moomau et al., 2016).

293

294 *Downstream sequence and structure affect PRF efficiency in a context-dependent manner*

295 The region downstream of the slippery site is thought to be crucial for PRF, typically because it creates
296 a roadblock for the translating ribosome by folding into a stable secondary structure like a hairpin or
297 a pseudoknot (Brierley and Pennell, 2001; Brierley et al., 1989; Parkin et al., 1992; Tu et al., 1992).
298 To get an estimate of the size of the relevant downstream region in different PRF events, we replaced
299 native regions with a constant sequence, leaving stretches of different length after the slippery site
300 unchanged (Fig 3A, top; Methods). While in the case of SARS almost the entire downstream region
301 included in our reporter construct (120 nt) was crucial for frameshifting, in the HIV and SIVmac239
302 PRF sites only a comparatively small region (<24 nt) was required to achieve PRF signal >50% of
303 wild type (Fig 3B). The shortness of the necessary downstream sequence in these cases was striking,
304 but in line with earlier evidence demonstrating that base pairing at the first 3-4 positions of the HIV
305 downstream stem-loop structure shows the strongest association with PRF efficiency (Mouzakis et
306 al., 2013).

307 To identify specific positions crucial for PRF we performed scanning mutagenesis of regions ranging
308 from 7 to 36 nt downstream of the slippery site (Fig 3A, bottom). This revealed large differences in
309 the fraction of relevant downstream positions (Fig 3C), ranging from 6% of single point mutations
310 reducing -1 PRF efficiency by more than half in the case of SIVmac239 to 53% in the case of SARS
311 (46% for +1 PRF at the OAZ1 site). The HIV gag-pol PRF site (among others) shows remarkable
312 resilience to point mutations (Fig 3C, Fig S6A), in line with the high degree of genetic variability in
313 HIV. Specific mutations affecting PRF typically disrupt base-pairing (e.g. positions 11, 12, 29 and 30
314 in HIV and positions 8-13 and 18-26 in SRV1; Fig S6A) and therefore change the (predicted)
315 secondary structure. In the case of HIV only mutations in the middle of the stem are detrimental,
316 suggesting that a reduced stem is still able to drive PRF at wild-type levels, and only forcing a
317 complete change in the structure affects frameshifting.

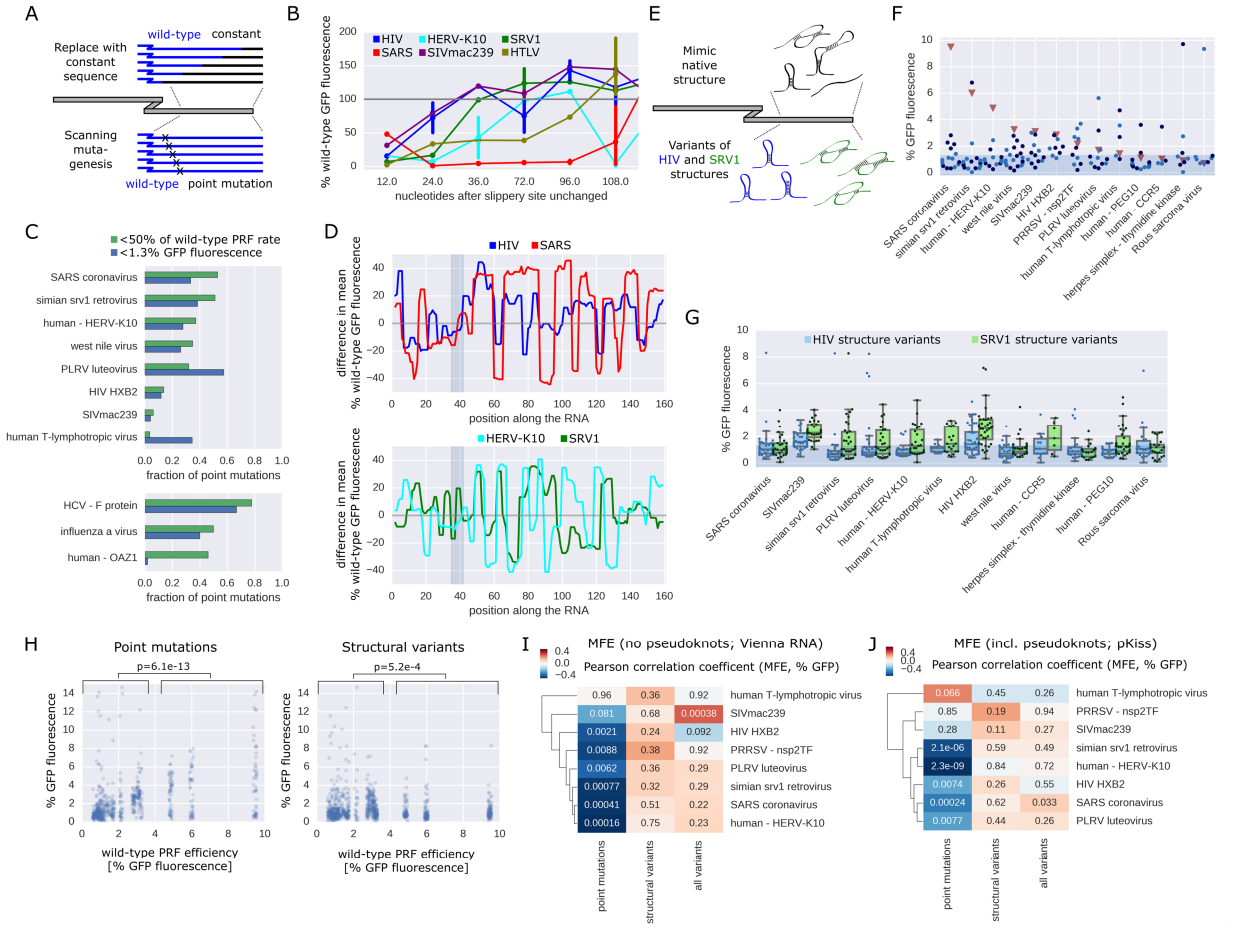

Figure 3. Downstream RNA structure affects PRF efficiency in a context-dependent manner.

A. Schematic of downstream sequence manipulations. B. Percent wild-type frameshifting rates of variants in which the native downstream region has been replaced by constant sequences, leaving the indicated number of nucleotides after the slippery site unchanged; n=1-3 per data point. C. Fraction of point mutations in the 40 nt downstream of the slippery site resulting in <50% of wild-type GFP fluorescence (green) and background fluorescence (<1.3% GFP fluorescence, blue), for -1 (top) and +1 (bottom) PRF events. D. The difference in mean % wild-type frameshifting between variants in which the indicated position is predicted to be paired vs unpaired along the variable region; gray box: slippery site. E. Schematic of downstream structure variations. F. Percent -1 GFP fluorescence of variants in which the downstream region is replaced by a synthetic sequence predicted to fold into the respective native secondary structure (using Vienna RNA (only stem-loop; light blue) or pKiss (including pseudoknots; dark blue; Methods)); the shaded area denotes the range of background fluorescence. G. % -1 GFP fluorescence of variants in which the downstream region is replaced by a synthetic sequence predicted to fold into the HIV (blue) or SRV1 (green) secondary structure (or variations thereof; Methods); the shaded area denotes the range of background fluorescence. H. % GFP fluorescence of variants in which the downstream region carried a single point mutation (left) or was replaced with sequences predicted to fold into variations of PRF inducing structures, plotted against the corresponding wild-type values; p-values (Mann-Whitney-U) for comparisons between groups of low (<4% GFP fluorescence) and high wild-type PRF (>4% GFP fluorescence). IJ. Clustered heatmap of Pearson correlation coefficients between percent -1 GFP fluorescence and minimum free energy of regions after the slippery site (computed using Vienna RNAfold on the 40 nt following the slippery site (I) or pKiss on the 120 nt following the slippery site (J)) for groups of variants of the indicated PRF sites; values denote the p-values of the correlations.

318

319 We tested if we can detect these mutational signatures of secondary structure also on the scale of
320 the entire library and compared the median PRF signal between all variants of a PRF event, in which
321 this position is predicted to be paired or unpaired (Fig 3D). Preferences for downstream positions to
322 be paired or unpaired reveal the architecture of structural elements downstream of the PRF site.
323 Groups of frameshifting events show remarkable concordance between these preferences (e.g. HIV
324 and SARS, HERV-K10 and SRV1), but length and position of the optimal downstream secondary
325 structures differ between groups (Fig 3D, upper vs. lower panel).

326 Replacing the endogenous downstream region with elements that have different primary sequence,
327 but are predicted to fold into the same secondary structure (Fig 3E, top) abolished frameshifting in
328 most cases, regardless of whether pseudoknots were included or not (Vienna RNA vs. pKiss to
329 determine the structure that served as the input for antaRNA (Kleinkauf et al., 2015a, 2015b); Fig 3F).
330 Introducing variants of the SRV1 and the HIV downstream structure (without preserving the original
331 sequence, Fig 3E, bottom; Methods) had the ability to trigger frameshifting in some cases (Fig 3G),
332 including ones where the wild-type sequence did not yield a frameshifting signal in the context of our
333 assay (CCR5 and PEG10). Strikingly, events associated with higher wild-type frameshifting rates
334 could not be "rescued" by introducing a secondary structure variant from a different PRF event (Fig
335 3H), indicating that more efficient frameshifting sites like SARS seem to be highly optimized, but less
336 tolerant to changes in the type of downstream secondary structure.

337 Changes in secondary structure (minimum free energy) induced by point mutations downstream of
338 the slippery site showed strong correlation with most, but not all frameshifting events (Fig 3I left
339 column, Fig S7 left panels). Surprisingly, this correlation between predicted minimum free energy of
340 secondary structure variants of the native downstream region and GFP fluorescence of -1 PRF
341 reporters was lost when replacing the downstream region with structural variants (more) different from
342 the native sequence (Fig 3I middle column, Fig S7 middle panels; these structures generally had the
343 potential to induce frameshifting in certain contexts; Fig 3G). Likewise when testing for correlation
344 across all variants of a PRF site (Fig 3I right column, Fig S7 right panels), there was no association
345 of lower minimum free energy (MFE) with higher PRF rates.

346 Many stimulatory downstream sequences are thought to fold into a pseudoknot (e.g. SRV-1 (ten Dam
347 et al., 1994; Du et al., 1997), west nile virus (Firth and Atkins, 2009) and coronaviruses (Brierley et
348 al., 1989; Plant et al., 2005)), a more complex folding pattern that is not covered by many secondary
349 structure predictions including the Vienna RNA package (Hofacker et al., 2004; Lorenz et al., 2011).
350 We therefore repeated the analyses using pKiss (Janssen and Giegerich, 2015), an algorithm that
351 can detect certain classes of pseudoknots (see Supplementary Data 2 for pKiss-predicted structures
352 of library variants with systematic sequence changes in the downstream region). MFE calculated
353 using Vienna RNAfold and pKiss, respectively, show good correlation (Fig S6B). We repeated the

354 above analysis and observed the same pattern as with RNAfold, namely widespread association
355 between secondary structure (low MFE) and higher PRF signal for small deviations from the native
356 sequence (Fig 3J left column, Fig S8 left panels), but not otherwise (Fig 3J middle and right column,
357 Fig S8 middle and right panels). This is in line with the fact that we do not see systematic differences
358 between PRF events whose downstream sequence was shown to fold into a hairpin structure (e.g.
359 HIV-1 (Parkin et al., 1992; Staple and Butcher, 2003) and SIVmac239 (Marcheschi et al., 2007)) or a
360 pseudoknot (e.g. SRV-1 (ten Dam et al., 1994; Du et al., 1997) and SARS (Brierley et al., 1989; Plant
361 et al., 2005)).

362 This observation does not imply a (counterintuitive) tendency for more open secondary structure to
363 promote frameshifting. As we do not find evidence for an effect of MFE on overall RNA levels (Fig S7,
364 S8, S9; e.g. through triggering No-Go decay (Belew et al., 2011)), we suggest that this reflects a
365 strong preference for downstream secondary structure rigidity to be close to that of the wild-type
366 sequence (in particular for SARS, SRV1 and HERV-K10, Fig S7, S8, right panels). Although present
367 secondary structure prediction algorithms might not be sufficiently accurate for this task, especially in
368 the case of pseudoknots, our results support a view according to which not only the structure, but also
369 the sequence downstream of a slippery site is critical for triggering frameshifting.

370

371 *Additional regulatory elements contribute to PRF*

372 The distance between the slippery site and the downstream secondary structure has been shown to
373 be important for efficient frameshifting (Brierley et al., 1989; Kollmus et al., 1994; Lin et al., 2012),
374 probably to position the paused ribosome at the slippery site (Brierley and Pennell, 2001). To test
375 whether this is generally true or if there are PRF event-specific differences, we introduced between
376 one and nine nucleotides from different constant sequences or deleted between one and six, mutating
377 downstream stop codons as required (Fig 4A top). Most insertions or deletions led to substantially
378 reduced frameshifting signal (Fig 4B), no matter what the context and if the original downstream
379 coding frame was preserved or not, suggesting that the exact position of the downstream regulatory
380 elements is important across PRF events. This is in contrast to an earlier study reporting unchanged
381 or even dramatically increased -1 frameshifting efficiency at the HIV site upon reducing spacer
382 lengths, which the authors attribute to a change in local secondary structure triggered by the specific
383 deletion introduced in their construct (Mouzakis et al., 2013). Overall, our results are in line with the
384 regulatory elements being in optimal distance from one another and the downstream secondary
385 structure creating a roadblock that causes the ribosome to pause exactly at the slippery site (Brierley
386 and Pennell, 2001; Tu et al., 1992).

387 While mutations in the slippery site and the downstream region typically have a negative effect on
388 frameshifting efficiency (Fig 2A), changes in the upstream region have the potential to lead to higher

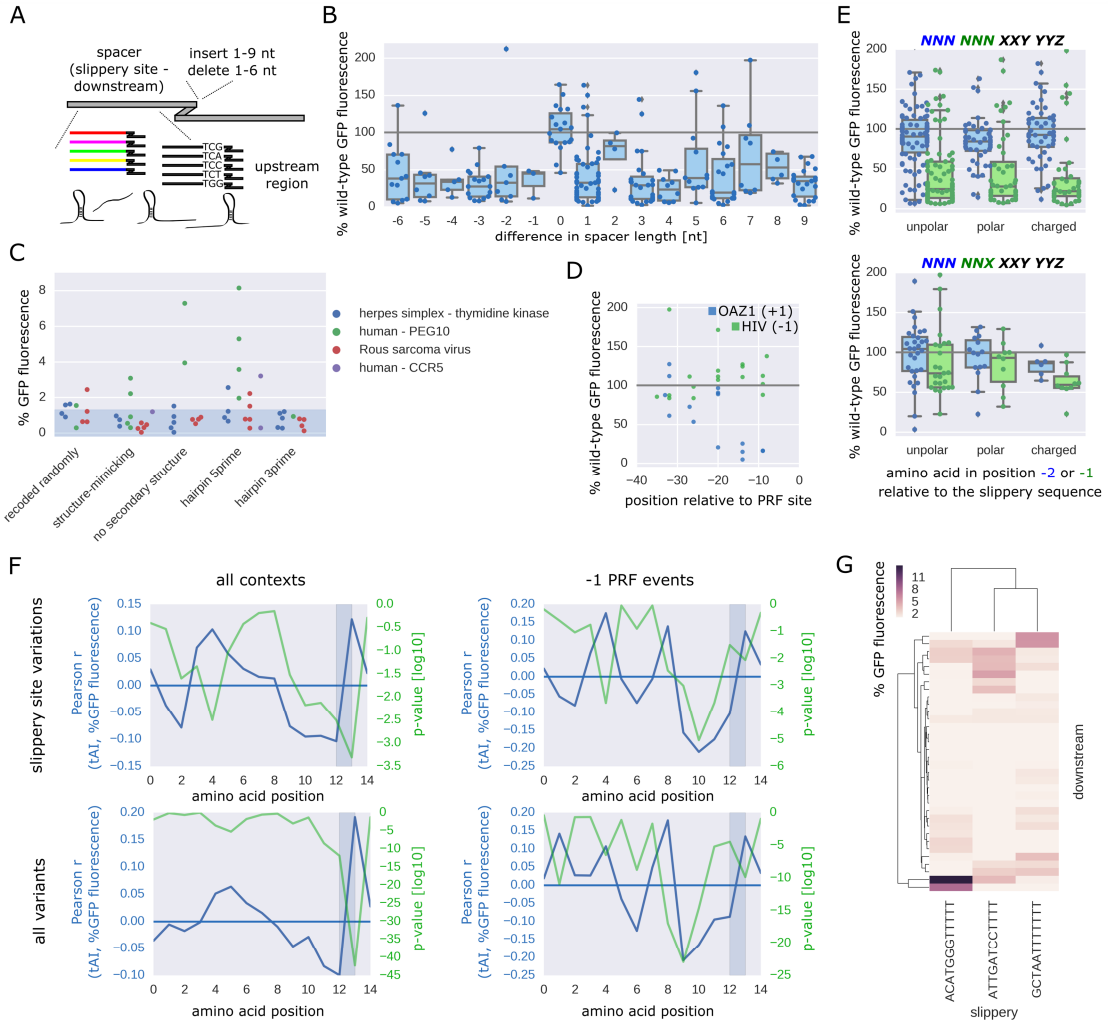

Figure 4. Additional regulatory elements contribute to PRF.

A. Schematic of sequence manipulations upstream and immediately downstream of the slippery site. B. Boxplots of % wild-type frameshifting rates for variants of “positive” -1 PRF sites in which the indicated number of nucleotides were deleted (negative numbers) or inserted (positive numbers, taken from two different constant sequences) after the slippery site (n=4-54). C. % GFP fluorescence for variants of “negative” -1 PRF sites with the sequence upstream of the slippery site being randomly recoded or replaced with sequences predicted to have the indicated secondary structure; the shaded area denotes the range of background fluorescence. D. % wild-type frameshifting rates (+1 for OAZ, blue, and -1 for HIV, green) of variants in which the native upstream region has been replaced by constant sequences up to the indicated position relative to the PRF site. E. Boxplot of percent of wild-type frameshifting rates for variants in which the -2 or -1 amino acid relative to the slippery site is replaced with an amino acid from the indicated groups, for codons which maintain the slippery site pattern XXXYYYZ (bottom) or not (top); n=10-115. F. Pearson correlation coefficient (blue) and associated p-values (green) between tAI at the indicated position of the original reading frame and % GFP fluorescence. G. Clustered heat map showing all possible combinations of 3 synthetic slippery sites, 34 synthetic downstream variants and up to 5 synthetic upstream regions (minimal value 1.3%).

-1 PRF signal (e.g. PEG10, Fig S5D). This suggests that inhibitory signals upstream of the frameshifting site as found for SARS (Su et al., 2005) could be a more widespread property of PRF sites, probably creating a balance to ensure that frameshift promoting signals like a rigid downstream

392 secondary structure do not lead to a complete inhibition of translation and potentially degradation of
393 the mRNA. To test for effects of upstream secondary structure on PRF, we replaced the region
394 upstream of the slippery site with a recoded version (encoding the same amino acid sequence), a
395 sequence predicted to fold into the same secondary structure as the wild-type, a sequence lacking
396 any strong predictable secondary structure or a hairpin at the 5' or 3' end. Specifically in the case of
397 PEG10, a hairpin at the 5' end of the upstream region or a lack of secondary structure resulted in a
398 frameshifting signal above threshold (2-8% GFP fluorescence; Fig 4C) as opposed to the wild-type
399 sequence (Fig 1C, see our discussion of potential reasons above).

400 The upstream region of OAZ1 is thought to harbor signals enhancing frameshifting (cf. Ivanov et al.,
401 2006). By testing variants in which increasing portions of the upstream region have been replaced
402 with constant sequences, we find these PRF-promoting signals to be located in the 20 nucleotides
403 before the slippery site (Fig 4D, decrease of PRF signal on average to 31.3% of wild-type, $p=0.0077$,
404 Wilcoxon signed-rank test, effect on region further upstream not significant). In contrast to the other -
405 1 PRF sites, HIV frameshifting rates show an increase upon mutation of the corresponding upstream
406 region (Fig 4D, mean increase of PRF signal to 128.2% of wild-type, $p=0.013$, Wilcoxon signed-rank
407 test, effect on region further upstream not significant), indicating the presence of an inhibitory element
408 in this region.

409 We expanded our search for properties affecting PRF efficiency and examined the effect of the codons
410 preceding the slippery site. We found that the presence of a charged amino acid immediately
411 upstream of the slippery site significantly reduced frameshifting signal, even when the sequence of
412 the slippery site was unchanged (Fig 4E, $p<0.0015$, Wilcoxon signed-rank test). We examined also
413 the influence of decoding efficiency as measured by the tRNA adaptation index (tAI) on PRF and
414 detected an increasingly negative correlation between tAI and PRF signal in the 0 frame codons
415 leading up to the site of ribosome slippage (Fig 4F). This observation might indicate that progressive
416 slowing down of the ribosome by limited tRNA availability might contribute to stalling at the slippery
417 site and frameshifting.

418 To not limit our analysis to endogenous frameshifting events, we designed completely synthetic
419 regions corresponding to a slippery site and the upstream and downstream region with different
420 sequence and structural properties. Some fully designed sequences were able to trigger frameshifting
421 and resulted in a -1 PRF signal of up to 11% GFP fluorescence, approximately corresponding to the
422 wild-type frameshifting rate of the most efficient event tested here (SARS coronavirus). We tested all
423 different downstream regions with synthetic slippery sequences resembling common types of -1 PRF
424 sites and found pronounced differences in combinatorial preferences (Fig 4G), showing that also fully
425 synthetic frameshifting events exhibit the combinatorial preferences and context-dependent
426 peculiarities observed for native PRF events.

427

428 *Prediction of frameshifting potential and efficiency*

429 Having accurate quantitative measurements for large collections of frameshifting sites, we aimed to
430 predict frameshifting efficiency using machine learning approaches. Based on previous studies and
431 our own findings, we used the identity of slippery site positions, tAI of codons around the frameshifting
432 site, amino acid class, minimum free energy and pairedness of positions downstream of the
433 frameshifting site (see Methods), alone and combined, as features and built computational models
434 based on Gradient Boosting Decision Trees (XGBoost (Chen and Guestrin, 2016); Fig 5A and
435 Methods). These models were either estimating whether a sequence has the ability to induce
436 frameshifting (i.e. yield a fluorescent signal above threshold (1.3%, see Figure S2A-C); classification)
437 or quantitatively predicting PRF efficiency (approximated by GFP fluorescence intensity in our assay;
438 regression). We achieved high accuracy (up to an area under the receiver operating characteristic
439 curve (ROC AUC)=0.93 for the classification of a sequence as yielding a PRF signal or not, and
440 Pearson r=0.81 for the comparison between measured and predicted frameshifting signal (% GFP
441 fluorescence)) when training our model on variants of specific frameshifting events and predicting
442 unseen variants from the same event (Fig 5BC). Despite the differences in the regulatory
443 characteristics between individual PRF events, we could predict frameshifting ability (Fig 5D, Fig
444 S10A, ROC AUC=0.89; area under the Precision-Recall-Curve=0.8) and frameshifting signal (% GFP
445 fluorescence; Fig 5E, Pearson r=0.53) also for a pool of all sequence variants tested, no matter
446 whether they yielded a detectable frameshifting signal in our assay or not (Fig 5BC, Fig S10A; "-1
447 PRF events" vs. "all library variants").

448 To determine to what extent a model trained on one PRF event could predict frameshifting potential
449 and -1 PRF signal in a different context, we trained our model (based on the full set of features) on
450 sequence variants derived from one event and predicted on variants of another event (Fig 5F, Fig
451 S10B). While performance is usually best for variants of the same PRF event, this approach revealed
452 functional similarities that allowed relatively accurate prediction on "foreign" PRF events (ROC AUC
453 up to 0.94, Pearson r up to 0.69). There is considerable overlap with other attempts of functional
454 clustering, e.g. based on sensitivities to mutations (Fig 2A), with SRV1 and HERV-K10 PRF sites
455 showing similar behavior and HIV variants yielding good prediction scores for models trained on data
456 from any retroviral PRF event (HIV, SIVmac239 and SRV1; Fig 5F). The SARS PRF event, however,
457 is not amenable to prediction based on models trained on data from other PRF events. Prediction
458 accuracy also reflects the specific components of particular importance for a PRF event, e.g. the
459 slippery site for HIV (Fig S10C) and the downstream region for SRV1 (Fig S10D). In these cases,

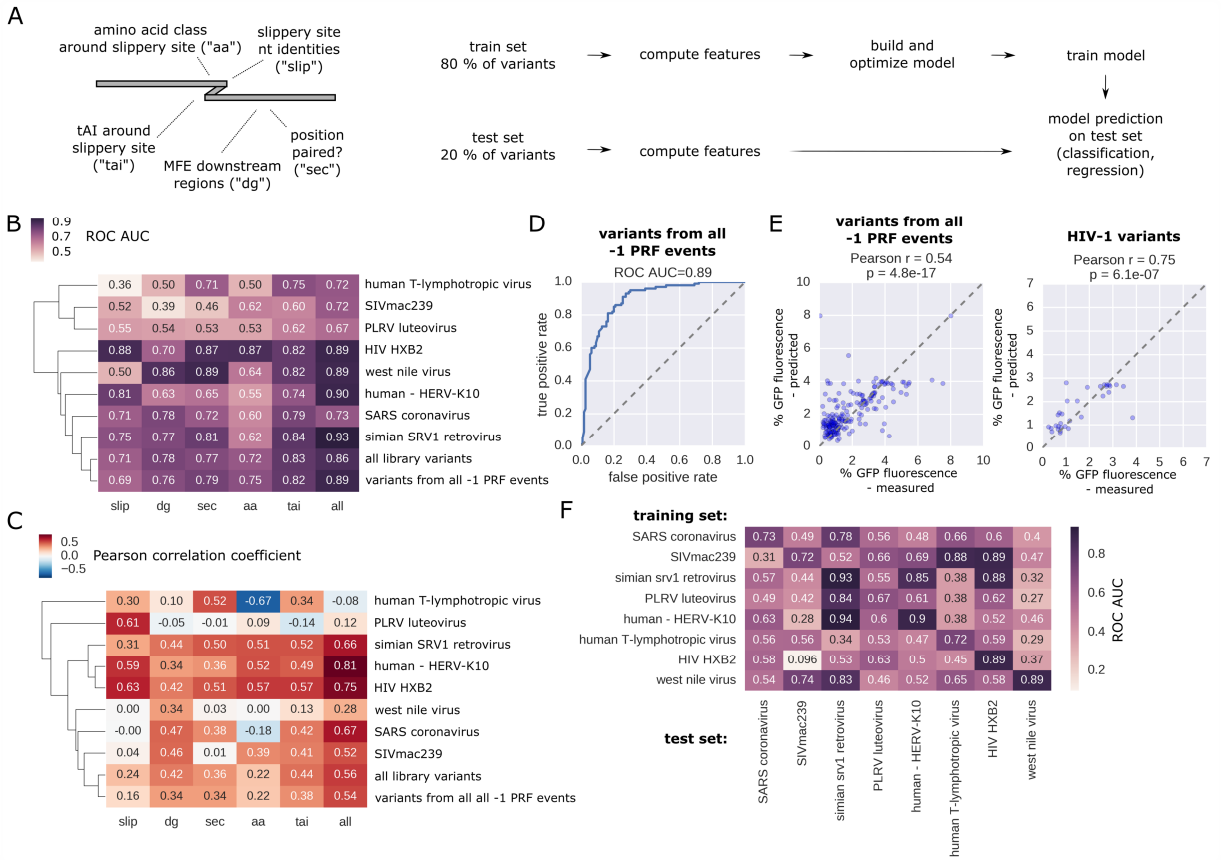

Figure 5. Prediction of frameshifting potential and efficiency.

A. Outline of the process of generating features and building and testing predictive models. BC. Prediction scores (area under the receiver operating characteristic curve ("ROC AUC") in the case of classification (B) and Pearson correlation coefficients in the case of regression (C)) on held-out test data for the indicated PRF event(s) or the entire library using different feature sets (slip: nucleotide identities in the canonical XXXYYYZ slippery site; tai: tAI scores around the slippery site, aa: amino acid class ("unipolar", "polar", "charged") around the slippery site, dg: MFE of downstream regions, sec: predicted pairedness of downstream positions). D. ROC curve showing performance of a classifier trained on 80% of all designed library variants passing filtering and testing on the remaining 20%. E. For the test set (20% of all designed library variants of all -1 PRF events (left) or HIV-1 (right) passing filtering) the measured % GFP fluorescence is plotted against the model prediction (trained on 80% of all library variants of all -1 PRF events (left) or HIV-1 (right) passing filtering). F. Prediction scores (ROC AUC) on held-out test data (20%) for the indicated PRF event (columns) using the full set of features and training the model on the training data (80%) from the indicated PRF event (rows).

frameshifting efficiency can be predicted using models trained on data from almost any other PRF event.

Interpretation of the prediction models provides an additional approach to identify the most important properties of a (potential) PRF site. Here, we used Shapely (SHAP) values for determining the contribution of each feature to the prediction result of every sample (Fig S11A; Lundberg et al., 2020). The features of the different subsets driving the prediction provide additional information about the

466 effect of specific sequence properties on the readout (Fig S11A), e.g. recapitulating earlier
467 observations that XXXCCCZ ("Y=C") and XXXGGGZ ("Y=G") slippery sites are disfavored (Fig S11A
468 upper left, cf. Fig 2D). For a model trained on the MFE of different lengths of downstream region,
469 stronger secondary structure not always led to prediction of higher PRF efficiency (Fig S11A lower
470 left). When taking into account regions of e.g. 60 nt after the slippery site ("dg60"), a lower feature
471 value favored prediction of PRF (Fig S11B).). While the positive effect of strong secondary structure
472 (low MFE) in the first 60 nt on PRF prediction was strongly enhanced if also the downstream region
473 as a whole showed low MFE (Fig S11B), in general we find strong secondary structure of the entire
474 downstream region (120 nt) to be associated with prediction of no PRF (Fig S11A). This could either
475 be explained through an effect on frameshifting efficiency or on RNA stability and overall expression
476 levels. The latter option is unlikely to explain the effect in full as we do not observe a correlation
477 between MFE of the downstream region and steady-state RNA levels (Fig S11C).

478

479 *HIV clinical isolates exhibit subtype-specific differences in frameshifting efficiency and associations*
480 *with viral load in patients*

481 The HIV gag-pol frameshifting site is arguably one of the most intensely studied examples of PRF
482 and – due to its critical importance for the viral replication cycle – has been repeatedly suggested as
483 an antiviral drug target (Brakier-Gingras et al., 2012; Hung et al., 1998). To assess the natural
484 variation in frameshifting rates in HIV1 we assembled a set of 581 sequences from clinical isolates
485 between 1976 and 2014 (http://www.hiv.lanl.gov/, Methods), which differ in the sequence surrounding
486 the frameshifting site, but not in the slippery site itself (Fig S12A). Frameshifting rates of the variants
487 are distributed around the rate observed for the lab strain HXB2 (Fig 6A). This constitutes actual
488 differences in frameshifting rates and not only experimental variability, as isolated clones show
489 remarkably good correlation given the small differences in PRF signal we are measuring (Fig 6B).
490 Secondary structure showed the best correlation with frameshifting rates when considering the first
491 30 nucleotides after the frameshifting site (Fig S12B), matching the region of high sequence
492 conservation (Fig S12A). We grouped the HIV variants based on subtype and found significant
493 differences between the groups (Fig 6C, $p<8*10^{-11}$, one-way ANOVA), most notably higher
494 frameshifting rates in subtype C ($p<6*10^{-12}$ for the difference between C and B), in contrast to an
495 earlier report (Baril et al., 2003). HIV subtypes show distinct geographical distributions (Hemelaar et
496 al., 2006), and consequently we also observed differences between countries of origin (Fig S12C,
497 $p<2*10^{-4}$), but no change in frameshifting rates over time (Fig S12D, p=0.4).

498 Optimal gag-pol frameshifting rates have been proposed to be critical for virulence (Dinman and
499 Wickner, 1992; Felsenstein and Goff, 1988). In order to link the frameshifting rate measured in our
500 reporter assay with the replicative success of the corresponding HIV isolate we compared the viral

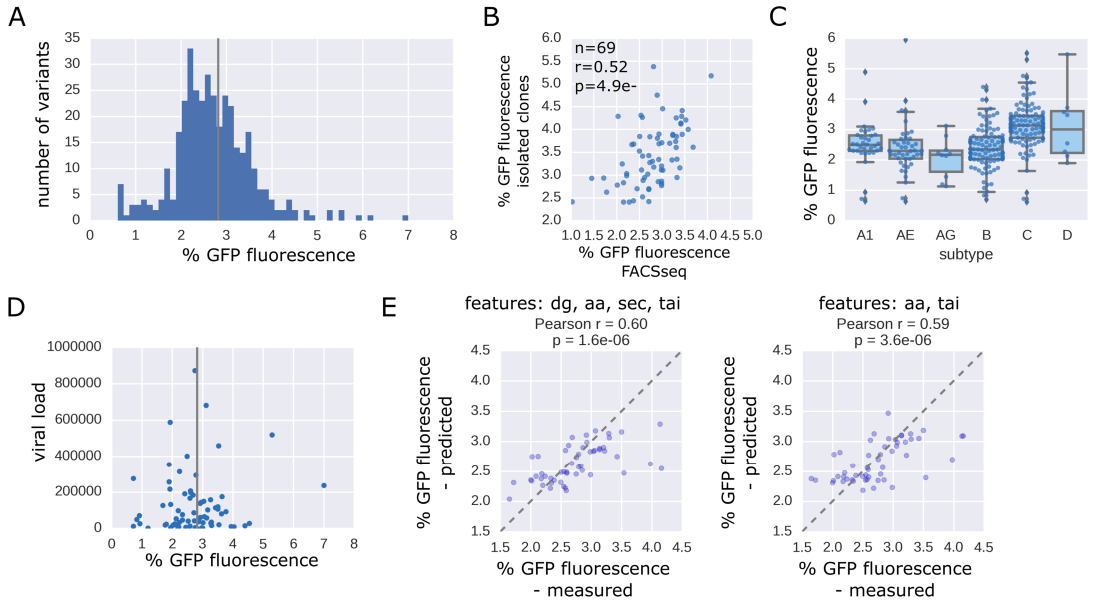

Figure 6. Testing of PRF sites from HIV clinical isolates reveals subtype specific differences and associations with viral load in patients.

A. Distribution of percent GFP fluorescence of all measured HIV variants. B. Percent GFP fluorescence as determined by FACSseq is plotted against the value determined by measuring the corresponding isolated clone by FACS. C. Boxplot showing percent GFP fluorescence of HIV gag-pol PRF variants coming from the indicated subtypes (n=8-112). D. Viral load (HIV titer) is plotted against percent GFP fluorescence determined for the natural gag-pol PRF site variant from the corresponding clinical isolate; vertical line: HIV HXB2 wild-type percent GFP fluorescence. E. Predicted vs measured percent GFP fluorescence for 20% of the HIV variants (held-out test set) with the full feature set except for the identity of slippery site positions (left) or after removing MFE features (middle) or all secondary structure derived features (right) features.

load in patients (where available) to frameshifting rates of the causal HIV isolate (Fig 6D). Although viral load is influenced by many factors, we nevertheless observed a clear trend for patients with high viral load to have frameshifting rates close to ”wild type” (the lab strain HXB2). These data present optimality of frameshifting rates as a hallmark of HIV infection and as being associated with infectious success, underscoring the potential for drugs altering the efficiency of gag-pol frameshifting.

In order to predict frameshifting efficiency of naturally occurring variants of the HIV gag-pol frameshifting site, we built a model based on the feature sets described above. Despite eliminating the most important feature of the HIV gag-pol PRF site (the slippery sequence) by restricting our analysis to clinical isolates with no sequence alterations in and around the canonical UUUUUUA and despite the narrow range of frameshifting rates (Fig 6A), we could accurately predict frameshifting rates of unseen variants (Pearson r=0.6; Fig 6E, left). Interestingly, removing the features based on pairedness of downstream positions and minimum free energy led to similar agreement between measured and predicted values (Fig 6E, right). This is in line with the limited importance of the exact downstream structure in comparison to other PRF events tested. Including designed HIV variants in the training set yielded similar prediction accuracy (Fig S12E), indicating that our model can learn the

516 relevant rules from our set of natural variants alone. Taken together, these results show that a model
517 based on data from our assay is sensitive enough to detect even subtle differences between HIV
518 variants and allows prediction of frameshifting rates with an accuracy of clinical relevance.

519

520 **Discussion**

521 Here, we combined fluorescent frameshifting reporters with rational design of DNA sequences and
522 high-throughput testing to systematically decipher the rules governing PRF. Our approach aims at a
523 systematic and comparative assessment of the commonalities and peculiarities in the regulation of
524 frameshifting efficiency across PRF events. It is complementary to in-depth analyses of individual
525 frameshifting sites yielding detailed structural or kinetic insights (Caliskan et al., 2014, 2017; Chen et
526 al., 2014; Choi et al., 2020; Lyon et al., 2019; Yan et al., 2015). The high-throughput nature of our
527 approach entails particular limitations. While we show that our assay is highly sensitive to changes in
528 the PRF signal (e.g. in the testing of natural isolates of HIV) and therefore able to identify even subtle
529 effects of sequence alterations on frameshifting, our detection limit is higher than in assays using
530 amplification methods and/or overexpression of the reporter.

531 Prediction of PRF events has up to now been largely limited to identifying sequences matching the
532 canonical pattern for slippery sites, followed by a downstream secondary structure (Belew et al., 2008;
533 Theis et al., 2008). To the best of our knowledge, no attempts at quantitatively predicting the effect of
534 sequence variation exist to date. We used our measurements of frameshifting signal of thousands of
535 variants of known PRF sites to build a computational model based on known and novel sequence
536 features affecting frameshifting, leading to accurate prediction of frameshifting potential (up to ROC
537 AUC=0.93) and frameshifting rates (up to Pearson r=0.81). In many cases sequence properties other
538 than slippery site identity or downstream secondary structure rigidity yielded high prediction scores,
539 emphasizing the importance to include additional features such as amino acid properties and tRNA
540 availability in the investigation of PRF regulation.

541 By controlled sequence and structure manipulations in multiple contexts we aimed to dissect and
542 directly compare the regulatory architecture of 15 frameshifting events, revealing a great diversity in
543 regulatory strategies involving upstream and downstream sequence and structural elements and
544 amino acid properties. We demonstrate a high degree of optimization and specialization of regulatory
545 mechanisms. For example, no universally optimal slippery site exists, but in each of the PRF events
546 tested the native slippery site was found to be optimal given the sequence context.

547 A correlation between secondary structure rigidity and PRF efficiency has been shown previously
548 (Bidou et al., 1997; Garcia-Miranda et al., 2016; Wu et al., 2018), while also other properties of
549 secondary structures like their plasticity have been reported to be associated with frameshifting rates

550 (Halma et al., 2019; Ritchie et al., 2012). Our results corroborate the functional link between minimum
551 free energy of downstream structures and show that this is true for most PRF events tested. However,
552 this correlation generally depended on the length of the downstream region taken into account, in
553 agreement with the specific association between PRF efficiency and thermodynamic stability of the
554 start of the HIV stem loop structure reported earlier (Mouzakis et al., 2013). Furthermore, the
555 correlation between PRF signal and MFE only held true when testing small variations of the native
556 sequence and structure; when sequences and structures different from the native one were tested,
557 this correlation was lost.

558 Slippery site and a downstream stimulatory signal are central – and therefore intensely studied –
559 regulatory elements present in all known PRF events. Here, we expanded the repertoire of regulatory
560 inputs and demonstrated that both tRNA availability as well as amino acid properties can affect PRF
561 efficiencies across native sequence contexts.

562 Ribosome stalling due to limited tRNA availability has the potential to lead to slippage of the ribosome
563 (Gallant and Lindsley, 1992; Weiss et al., 1988) and contributes to non-programmed frameshifting in
564 huntingtin (Girstmair et al., 2013). Availability of tRNAs decoding slippery site codons have been
565 shown to influence HIV gag-pol frameshifting efficiency (Korniy et al., 2019) and HIV-1 itself has the
566 ability to modulate the tRNA pool of the host cell (van Weringh et al., 2011). Our observation that
567 limited tRNA availability in the codons leading up to the slippery site progressively contributes to PRF
568 efficiency is in line with these data and expands the notion of an influence of the tRNA pool on
569 frameshifting beyond the slippery site codons and individual PRF events.

570 In addition, our analyses highlighted amino acid properties around the frameshifting sites as another
571 factor influencing PRF efficiency. Specifically, a charged amino acid at the last position before the
572 slippery site was associated with reduced frameshifting rate. Together with reports of a frameshift-
573 promoting effect of specific amino-acid sequences upstream of the frameshifting site in the case of
574 +1 PRF in fungal OAZ genes (Yordanova et al., 2015) and the copA gene of Escherichia coli (Meydan
575 et al., 2017), this observation provides additional evidence for a role of the nascent peptide in
576 regulating ribosomal frameshifting.

577 In general, our results reveal functional groups of PRF events that exhibit greater similarity in terms
578 of their sensitivities to mutations in regulatory regions and congruence in the preferences for
579 secondary structure. These functional groups also manifest themselves in a higher prediction
580 accuracy within these groups. The groups resemble similarities also according to other criteria, e.g.
581 in their native secondary structure. PRF events, whose downstream stimulatory sequence has been
582 shown to be a stem loop (e.g. HIV-1 (Parkin et al., 1992; Staple and Butcher, 2003) and SIVmac239
583 (Marcheschi et al., 2007)) tend to be more resilient and less affected by small sequence changes than
584 those with a downstream pseudoknot (e.g. SRV-1 (ten Dam et al., 1994; Du et al., 1997) as well as

585 SARS and other coronaviruses (Brierley et al., 1989; Plant et al., 2005)). Typically, these cases exhibit
586 higher fluorescence signal in our assay (Fig 1C), but greater sensitivities to even small sequence
587 changes (Fig 2B, 3C, S4AB, S5A) and requirements for a more extensive downstream region. This
588 might also entail a greater sensitivity to the experimental context and therefore explain the lack of
589 signal (above the noise threshold) for previously reported cases depending on a pseudoknot as
590 frameshift stimulatory signal like Rous sarcoma virus (Jacks and Varmus, 1985; Marczinke et al.,
591 1998). In other cases, low wild-type rates of frameshifting, requirements for additional trans-acting
592 factors or an effect on overall expression levels might explain the lack of signal in our assay.

593 Comparing the HIV gag-pol site with other PRF events shows the remarkable robustness and
594 exceptional tolerance for various non-native downstream stimulatory signals. More than in other
595 cases, maintaining the wild-type slippery site UUUUUUA is crucial for frameshifting, and tends to be
596 absolutely invariable across otherwise diverse natural isolates (Biswas et al., 2004). The quantitative
597 nature of our assay allows us nevertheless to accurately quantify the effect of sequence variation
598 away from the slippery site, revealing sub-type specific differences. Furthermore, we devised a
599 machine learning model that is able to predict the natural variation in HIV frameshifting rates for novel
600 variants (Pearson r=0.60). Precise stoichiometry of frameshifted and non-frameshifted product is
601 crucial for HIV virulence, and the ability to predict frameshifting rates of naturally occurring variants of
602 the gag-pol PRF site therefore has important clinical implications, both in the case of HIV as well as
603 in the case of PRF events in other viruses.

604 This high-throughput investigation of PRF also constitutes an addition to the experimental toolbox for
605 studying translational recoding and highlights many starting points for further investigations into the
606 regulatory basis and functional importance of frameshifting. In addition to deciphering the effect of cis-
607 regulatory elements, regulation by trans-acting factors like the recently discovered case of the
608 frameshift repressor shiftless (Wang et al., 2019) constitutes another area where a systematic
609 approach like the one presented here can yield novel insights into the mode of action and its
610 specificity. Libraries of frameshifting reporters provide a platform for screening putative modifiers of
611 ribosomal frameshifting, offering a powerful tool to identify ways to generally or selectively control
612 frameshifting and opening new possibilities for interfering with viral replication and for controlling
613 cellular processes depending on translational frameshifting.

614

615

616 **Methods**

617 **Synthetic library design**

618 *General design notes*

619 Oligonucleotides were designed to maintain a constant length of 210 nt. Restriction sites used for
620 cloning were excluded from the design. All the variants were composed of an 18 nt forward primer,
621 12 nt barcode sequence, 162 nt variable region and 18 nt reverse primer sequences. DNA barcodes
622 were designed to differ from any other barcode in the library in at least 3 nt.

623

624 *Selection of frameshifting sites for systematic sequence manipulations*

625 60 PRF events reported previously or listed in FSDB (Moon et al., 2007) were included in the library
626 design. Sequences were aligned in a way that the site of frameshifting would be at the same position
627 in all variants (slippery site ending at position 42). In case there were stop codons downstream of the
628 PRF site in any of the three frames, additional versions with stop codons mutated (changed to TGG)
629 were included for testing only. Out of these, 15 were selected for systematic sequence manipulations
630 based on the amount of experimental evidence supporting them, lack of stop codons in the relevant
631 reading frames, biological significance and mechanistic diversity (i.e. covering +1, -1 and -2
632 frameshifting sites). The HIV and OAZ frameshifting sites were tested beforehand in isolation to
633 confirm that our experimental setup is able to reproduce well established cases of -1 as well as +1
634 PRF.

635

636 *Design of individual rationally designed subsets*

637 For each of the subsets in the libraries, systematic sequence manipulations were performed on the
638 set of 15 previously reported frameshifting sites. Design of subsets was carried out in Python.

639 Multiple barcode controls: We added multiple variants to the library that contained the same variable
640 region, but different barcodes, in order to gauge potential effects of the barcode and the technical
641 noise of our assay.

642 Slippery site: We tested all possible variations of the different slippery site patterns found in previously
643 reported frameshifting events (XXXYYYZ, XYYZZZZ, influenza, OAZ (for +1 PRF)). In the case of
644 OAZ, both the first and the second codon in the slippery site UCC UGA were replaced with all other
645 codons. In addition, we created all possible point mutations in -1 PRF slippery sites that would not
646 give rise to a stop codon.

647 Immediate downstream region/"spacer": We created all possible point mutations in the six nucleotides
648 downstream of PRF slippery sites that would not give rise to a stop codon. In addition, we removed
649 up to 6 bps immediately downstream (in 1 or 3 bp increments, depending on the presence of stop
650 codons in downstream reading frames) or introduced up to 12 bps from five different constant
651 sequences (in 1 or 3 bp increments, depending on the presence of stop codons in downstream
652 reading frames).

653 Immediate upstream region: The first two codons preceding the slippery site (in the original frame)
654 were replaced with all other codons (except stop codons).

655 Upstream region: The upstream region (excluding the last two codons before the slippery site) was
656 randomly recoded five times. Moreover, variants were added for all 15 PRF events in which the same
657 part of the upstream region was replaced by four different constant sequences in increments of six
658 bps, starting from the beginning of the variable region. In addition, we included the following structural
659 variations in the design: We designed sequences 30 nt in length predicted (using ViennaRNA 2.0
660 (Lorenz et al., 2011) and antaRNA (Kleinkauf et al., 2015a, 2015b)) to fold into the respective native
661 structure (but having a different primary sequence or to have no secondary structure, a 9 bp long
662 stem-loop at the 5' or at the 3' end (6 sequences per set) and replaced the endogenous region in all
663 15 PRF events.

664 Downstream region: Variants were added for all 15 PRF events in which the same part of the
665 downstream region was replaced by three different constant sequences (different sets were used for
666 predicted +1 and -1 PRF sites to avoid stop codons in the relevant frames) in increments of six bps,
667 starting from the end of the variable region. For scanning mutagenesis, 30 positions (leaving the first
668 6 bp after the slippery site unchanged) were replaced with either A or C (G and T were avoided as
669 they often give rise to stop codons in one of the frames). In addition, we included the following
670 structural variations in the design: We used ViennaRNA 2.0 or pKiss (Janssen and Giegerich, 2015)
671 (which also predicts pseudoknots) to determine the native secondary structure of each of the 15 PRF
672 events. We then created for each PRF event two sets of ten sequences each (using antaRNA) that
673 are predicted to fold into the downstream secondary structure as determined using ViennaRNA and
674 pKiss, respectively. Furthermore we designed variants of the predicted HIV, SRV1 and nsp2F
675 (PRRSV) downstream secondary structure, e.g. by extending or reducing the length of hairpins, and
676 completely synthetic secondary structures and again used antaRNA to create sets of corresponding
677 sequences that we used to replace the native downstream structure in all 15 PRF events.

678 Completely synthetic variants: We designed a set of four upstream regions, three potential slippery
679 sites (resembling previously reported cases of the patterns XXXYYYZ, XYYZZZZ and XXYYYYY),
680 and created all possible combinations with a set of 57 artificial downstream sequences.

681 Frameshifts: We created variants in which the slippery site was moved by one, two or three positions
682 in the 5' or 3' direction in order to verify the frame-dependence of the effect.

683 Combinatorial variants: Upstream, slippery site and downstream regions from the whole set of 15
684 PRF sites were recombined, either preserving the original sequence fully or replacing potential stop
685 codons in all frames with TGG.

686

687 *gag-pol frameshifting sites from HIV clinical isolates*

688 We retrieved >30,000 full length sequences of the HIV genome from clinical isolates
689 (http://www.hiv.lanl.gov/) and selected gag-pol frameshifting sites which had the canonical sequence
690 around the slippery site (TAATTTTTTA), but differed in at least one position in the 30 nucleotides
691 upstream or the 120 nucleotides downstream. We chose to exclude variability in the slippery site to
692 screen for more subtle differences in frameshifting rates across clinical isolates.

693

694 **Experimental procedures**

695 *K562 cell culture*

696 K562 cells were acquired from ATCC. Cells were grown in Iscove's Modified Dulbecco Medium
697 supplemented with 10% fetal bovine serum (SIGMA) and 1% Penicillin-Streptomycin solution
698 (SIGMA). The cells were split when reaching a concentration of ~10^6 cells/ml. The cells were grown
699 in an incubator at 37°C and 5% CO^2. Cells were frozen in batches of 4×10^6 cells in growth medium
700 supplemented with 5% DMSO.

701

702 *Construction of the master plasmid*

703 Master plasmids for library insertion were constructed by amplifying parts from existing vectors and
704 cloning the parts sequentially into pZDonor 3.1. The master plasmid contained the EF1alpha
705 promoter, *mCherry*, a designed multiple cloning site containing restriction sites for library cloning
706 (RsrII and AscI), *gfp* and the SV40 terminator sequence. In the case of the master plasmid for cloning
707 the stop-vector library, two complementary DNA oligonucleotides (GACTGATAGCTGACTAGTCG
708 and GTCCGACTAGTCAGCTATCAG) were synthesized (IDT), annealed and introduced into the
709 vector described above (cut with RsrII) such that the mCherry-proximal RsrII site was destroyed and
710 the downstream RsrII site (after the inserted stop cassette) was subsequently used for inserting the
711 library.

712

713 *Synthetic library cloning*

714 The cloning steps were performed essentially as described previously (Vainberg Slutskin et al., 2018).
715 We used Agilent oligo library synthesis technology to produce a pool of 17,809 different fully designed
716 single-stranded 210-oligomers (Agilent Technologies, Santa Clara, CA), which was provided as a
717 single pool of oligonucleotides (10 pmol). The two subsets of this pool corresponding to the designed
718 and native libraries tested here were defined by unique amplification primers. The reverse primer
719 contained a spacer of either 4 or 5 nucleotides to create the offset in the *gfp* reading frame. The oligo
720 pool was dissolved in 200 µl Tris-ethylene-diamine-tetraacetic acid (Tris-EDTA) and then diluted 1:50
721 with Tris-EDTA, which was used as template for PCR. We amplified each of the four libraries by
722 performing 8 PCR reactions, each of which contained 19 µl of water, 5 µl of DNA, 10 µl of 5× Herculase
723 II reaction buffer, 5 µl of 2.5 mM deoxynucleotide triphosphate (dNTPs) each, 5 µl of 10 µM forward
724 primer, 5 µl of 10 µM reverse primer, and 1 µl Herculase II fusion DNA polymerase (Agilent
725 Technologies). The parameters for PCR were 95°C for 1 min, 14 cycles of 95°C for 20 s, and 68°C
726 for 1 min, each, and finally one cycle of 68°C for 4 min. The oligonucleotides were amplified using
727 library-specific common primers in the length of 35 nt, which have 18-nt complementary sequence to
728 the single-stranded 210-mers and a tail of 17 nt containing RsrII (forward primer) and AscI (reverse
729 primer) restriction sites. The PCR products were concentrated using Amicon Ultra, 0.5 ml 30K
730 centrifugal filters (Merck Millipore). The concentrated DNA was then purified using a PCR mini-elute
731 purification kit (Qiagen) according to the manufacturer's protocol. Purified library DNA (540 ng total)
732 was cut with the unique restriction enzymes RsrII and AscI (Fermentas FastDigest) for 2 hours at
733 37°C in two 40-µl reactions containing 4 µl fast digest (FD) buffer, 1 µl RsrII enzyme, 1 µl AscI enzyme,
734 18 µl DNA (15 ng/µl) and 16 µl water, followed by heat inactivation for 20 min at 65°C. Digested DNA
735 was separated from smaller fragments and uncut PCR products by electrophoresis on a 2.5% agarose
736 gel stained with GelStar (Cambrex Bio Science Rockland). Fragments were cut from the gel and
737 eluted using electroelution Midi GeBAflex tubes (GeBA, Kfar Hanagid, Israel). Eluted DNA was
738 precipitated using sodium acetate–isopropanol. The master plasmid was cut with RsrII and AscI
739 (Fermentas FastDigest) in a reaction mixture containing 6 µl FD buffer, 3 µl of each enzyme and 3.5
740 µg of the plasmid in a total volume of 60 µl. After incubation for 2.5 hours at 37°C, 3 µl FD buffer, 3 µl
741 alkaline phosphatase (Fermentas) and 24 µl water were added and the reactions were incubated for
742 an additional 30 mins at 37°C followed by 20 min at 65°C. Digested DNA was purified using a PCR
743 purification kit (Qiagen). The digested plasmid and DNA library were ligated for 30 min at room
744 temperature in a 10 µl reactions, containing 150 ng plasmid and the library in a molar ratio of 1:1, 1
745 µl CloneDirect 10× ligation buffer, and 1 µl CloneSmart DNA ligase (Lucigen Corporation), followed
746 by heat inactivation for 15 min at 70°C. Ligated DNA was transformed into *E. coli* 10G
747 electrocompetent cells (Lucigen) divided into aliquots (23 µl each, plus 2 µl of the ligation mix), which
748 were then plated on 4 Luria broth (LB) agar (200 mg/ml amp) 15-cm plates per transformation reaction

749 (25 µl). The rationally designed (12,809 variants) and the native (5000 variants) parts of the library
750 were cloned separately. For the rationally designed and the native library we collected around 10^6 and
751 $4x10^5$ colonies, respectively, for each frame the day after transformation by scraping the plates into
752 LB medium. Library-pooled plasmids were purified using a NucleoBond Xtra maxi kit (Macherey
753 Nagel). To ensure that the collected plasmids contain only a single insert of the right size, we
754 performed colony PCR (at least 48 random colonies per library).

755

756 *Transfection into K562 cells and genomic integration*

757 The purified plasmid libraries were transfected into K562 cells and genomically integrated using the
758 Zinc Finger Nuclease (ZFN) system for site-specific integration and the CompoZr® Targeted
759 Integration Kit - AAVS1 (SIGMA). Transfections were carried out using Amaxa® Cell Line
760 Nucleofector® Kit V (LONZA). To ensure library representation we performed 15 nucleofections of
761 the purified plasmid library for each rationally designed library (frame -1 and frame +1) and 5 for each
762 native library. For each nucleofection, 4×10^6 cells were centrifuged and washed twice with 20 ml of
763 Hank's Balanced Salt Solution (HBSS, SIGMA). Cells were resuspended in 100 µl solution (warmed
764 to room temperature) composed of 82 µl solution V and 19 µl supplement (Amaxa® Cell Line
765 Nucleofector® Kit V). Next, the cells were mixed with 2.75 µg of donor plasmid and 0.6 µg ZFN mRNA
766 (prepared in-house) just prior to transfection. Nucleofection was carried out using program T-16 on
767 the NucleofectorTM device, immediately mixed with ~0.5 ml of pre-cultured growth medium and
768 transferred to a 6 well plate with additional 1.5 ml of pre-cultured growth medium. A purified plasmid
769 library was also transfected without the addition of ZFN and served as a control to determine when
770 cells lost non-integrated plasmids.

771

772 *Sorting the library by FACS*

773 K562 cells were grown for at least 14 days to ensure that non-integrated plasmid DNA was eliminated.
774 A day prior to sorting, cells were split to ~0.25×106 cells/ml. On the day of sorting, cells were
775 centrifuged, resuspended in sterile PBS and filtered using cell-strainer capped tubes (Becton
776 Dickinson (BD) Falcon). Sorting was performed with BD FACSAria II SORP (special-order research
777 product) at low sample flow rate and a sorting speed of ~18000 cells/s. To sort cells that integrated
778 the reporter construct successfully and in a single copy (~4% of the population), we determined a
779 gate according to mCherry fluorescence so that only mCherry-expressing cells corresponding to a
780 single copy of the construct were sorted (mCherry single population). We collected around 350
781 cells/variant on average for each library in order to ensure adequate library representation.

782 Cells sorted for single integration of the transgene were grown for a week before we sorted the
783 population (combined rationally designed and native library, but separately for -1 PRF reporters, +1
784 PRF reporters and the stop-vector control) into 16 (in the case of the stop-vector 12) bins according
785 to GFP fluorescence, after gating for a narrow range of mCherry expression to avoid effects coming
786 from the influence of the variable region on overall expression level. This resulted in the loss of a
787 significant fraction of variants (as only around 50% of the cells fell into the narrow range of mCherry
788 fluorescence) but ensured that the subsequent sorting by GFP fluorescence was not biased by
789 expression level. Each bin was defined to span a range of GFP fluorescence values such that it
790 contains between 1-10% of the cell population. We collected around 1000 cells/variant on average in
791 order to ensure adequate library representation. Cells from each bin were grown separately for
792 freezing and purification of genomic DNA.

793

794 *Genomic DNA purification, amplification and sample preparation*

795 For each of the bins we purified genomic DNA by centrifuging 5×10^6 cells, washing them with 1 ml
796 PBS and purifying DNA using DNeasy Blood & Tissue Kit (Qiagen) according to the manufacturer's
797 protocol. In order to maintain the complexity of the library amplified from gDNA, PCR reactions were
798 carried out on a gDNA amount calculated to contain a minimum average of 200 copies of each oligo
799 included in the sample. For each of the 16 bins, we used 15 µg of gDNA as template in a two-step
800 nested PCR. In the first step, three reactions were performed and each reaction contained 5 µg gDNA,
801 25 µl Kapa Hifi ready mix X2 (KAPA Biosystems), 2.5 µl 10 µM 5′ primer, and 2.5 µl of 10 µM 3′ primer.
802 The parameters for the first PCR were 95°C for 5 min, 18 cycles of 94°C for 30s, 65°C for 30s, and
803 72°C for 60s, each, and one cycle of 72°C for 5min. In the second PCR step, each reaction contained
804 2.5 µl of the first PCR product, 25 µl of Kapa Hifi ready mix X2 (KAPA Biosystems), 2.5 µl 10 µM 5′
805 primer, and 2.5 µl 10 µM 3′ primer. The PCR program was 95°C for 5 min, 24 cycles of 94°C for 30s
806 and 72°C for 30s, each, and one cycle of 72°C for 5min. Specific primers corresponding to the constant
807 region of the plasmid were used. The 5′ primer contained a unique upstream 8-nt bin barcode
808 sequence, and three different barcodes were used for each bin. The 3′ primer was common to all bins.
809 Multiple PCR reaction products of each bin were combined. The concentration of the PCR samples
810 was measured using a monochromator (Tecan i-control), and the samples were mixed in ratios
811 corresponding to their ratio in the population, as defined when sorting the cells into the 16 bins.
812 Sample preparation including gel elution and purification were performed as described above for
813 amplicons from cDNA.

814

815 *RNA purification, cDNA synthesis and sample preparation*

816 For the cell population sorted for single integration of the reporter construct we performed RNA
817 purification by centrifuging 10^7 cells, washing them with PBS, splitting into two tubes and purifying
818 RNA using NucleoSpin RNA II kit (MACHEREY-NAGEL) according to the manufacturer's protocol.
819 We prepared cDNA in 2 reverse transcription reaction for each replicate using SuperScript® III First-
820 Strand Synthesis System (Thermo Fisher Scientific) with random hexamer primers and 5 µg of input
821 RNA (per reaction) according to the manufacturer protocol. For amplification of the library variants, 3
822 PCR reactions of 50 ul total volume were performed. Each reaction contained 5 ul cDNA, 25 µl of
823 Kapa Hifi ready mix X2 (KAPA Biosystems), 2.5 µl 10 µM 5′ primer, and 2.5 µl 10 µM 3′ primer. The
824 PCR program was 95◦C for 5 min, 20 cycles of 94◦C for 30s and 72◦C for 30s, each, and one cycle
825 of 72◦C for 5min. Specific primers corresponding to the constant region upstream and downstream of
826 the splice sites were used. The PCR products were separated from potential unspecific fragments by
827 electrophoresis on a 1.5% agarose gel stained with EtBr, cut from the gel, and cleaned in 2 steps: gel
828 extraction kit (Qiagen) and SPRI beads (Agencourt AMPure XP). The sample was assessed for size
829 and purity at the Tapestation, using high sensitivity D1K screenTape (Agilent Technologies, Santa
830 Clara, California). We used 20 ng library DNA for library preparation for NGS; specific Illumina
831 adaptors were added, and DNA was amplified using 14 amplification cycles. The sample was
832 reanalyzed using Tapestation.

833 For RT-PCR analysis we performed RNA purification on 5 million cells from individually constructed
834 and integrated reporter constructs, followed by cDNA synthesis as described above (one reaction).
835 PCR was performed using primers within the mCherry and GFP coding region, yielding a PCR product
836 of about 1000 bp length. 40 cycles of PCR were used to obtain a strong signal even for less prominent
837 bands. PCR products were analyzed on a 1% agarose gel.

838

839 **Computational analyses**

840 *Mapping next generation sequencing reads*

841 To unambiguously identify the variant of origin, a unique 12-mer barcode sequence was placed at the
842 5' end of each variable region. DNA was sequenced on a NextSeq-500 sequencer. We determined
843 for each read (PE150) its bin barcode and its variant barcode and discarded all the reads that could
844 not be assigned to a bin and a library variant of origin. In a second step, for each variant the mapped
845 reads (paired-end, 150 bases for each mate) were aligned to the designed sequence (in the context
846 of the reporter) and only those ones were counted that mapped without a mismatch over the whole
847 length (disregarding the last 5 bases). Reads showing a mismatch, in particular a single nucleotide
848 deletion or insertion, were quantified separately and if they passed the same threshold applied to
849 correct library variants a PRF readout (% GFP fluorescence) was calculated for them. These

850 additional variants were only used in comparisons with the original designed variant and not
851 incorporated in the general pool of library variants considered for the analyses. For all three libraries
852 (-1 PRF reporters, +1 PRF reporters and the stop-vector control) around 40 milllion successfully
853 mapped reads were used for subsequent analysis. In addition, we sequenced the entire cell
854 population using greater read lengths (PE300) to cover the entire length of the variable region and to
855 obtain a measure of the fraction of reads per variant containing a synthesis or cloning error. Mapping
856 was performed using custom-made Python scripts.

857

858 *Computing a measure of frameshifting efficiency (% GFP fluorescence) and related readouts*

859 We applied a number of filters to the raw sequencing data to reduce experimental noise. First, variants
860 with less than 20 reads (perfect matches along the whole length) mapped across bins were removed.
861 Second, for bins with a read count of less than three, the bin value was set to zero. Third, for each
862 variant we set to zero bins surrounded by zero values as these constituted isolated bins unlikely to
863 come from the actual distribution. Forth, to reduce bias coming from the "open end"-bins at the
864 extreme values, we set their count to match their neighbor's if it was higher, as these bins are defined
865 as containing the tails of the distribution of variants with no GFP fluorescence and maximal GFP
866 expression, respectively, and peaks in these extreme bins are considered experimental noise. For
867 each variant, we normalized the values across the 16 bins by dividing the filtered read counts for each
868 bin by the sum of all filtered read counts. To determine whether a distribution is unimodal we applied
869 a Savitzky-Golay filter for smoothing the data and removing local maxima introduced by experimental
870 noise. We detected peaks in the smoothed distribution by a simple approach in which a point is
871 considered a maximum peak if it has the maximal value, and was preceded by a value lower by 5%.
872 Variants without detectable peaks (due to the read number not reaching the threshold) were excluded
873 at this stage. Variants with more than one peak were filtered out at the level of analysis. For each bin,
874 we calculated the median of the log2 of GFP fluorescence intensity as measured by FACS for all the
875 cells sorted into that bin and used this as the GFP fluorescence value associated with this bin. For
876 each variant with one peak after smoothing, we calculated the weighted average for the distribution
877 of reads across bins (using unsmoothed read counts normalized for each variant and the GFP
878 fluorescence value for the bins, as described above), resulting in what is referred to in the main text
879 and figures as the "GFP fluorescence" (in log2).

880 As described in the main text and illustrated in Figure S2A-C, we set the lowest value obtained (the
881 variant with the lowest weighted average of the distribution) to 0% GFP fluorescence and the highest
882 value obtained (the variant with the highest weighted average of the distribution) to 100% GFP
883 fluorescence. Accordingly, a percentage was assigned to each variant based on the GFP
884 fluorescence (log2 of the weighted average of the reads distribution) of a variant as follows:

885 $[\%\ GFP\ fluorescence]_{variant\ X} = ([GFP\ fluorescence]_{variant\ X} - min([GFP\ fluorescence]_{all\ variants})) /$
886 $(max([GFP\ fluorescence]_{all\ variants}) - min([GFP\ fluorescence]_{all\ variants})) * 100$

887 "% wild-type GFP fluorescence" of a variant was calculated as follows:

888 $[\%\ wild\text{-}type\ GFP\ fluorescence]_{variant\ X} = [\%\ GFP\ fluorescence]_{variant\ X} / [\%\ GFP\ fluorescence]_{wild\text{-}type\ version}$
889 $_{of\ variant\ X} * 100$As a measure for steady state RNA expression levels, we computed the log ratio of
890 RNA/DNA reads for all variants with at least 20 DNA reads past filtering: $log2([RNA\ reads]_{variant\ X} /$
891 $[DNA\ reads]_{variant\ X})$.

892

893 *Machine learning approaches*

894 Machine learning procedures were carried out using the python scikit-learn (version 0.18.2) and
895 XGBoost package. Initially, from all duplicated sequences (e.g. barcode control sets), which passed
896 filtering, a single variant was randomly chosen for all subsequent steps to avoid biases resulting from
897 having duplicated sequences. We created sets of variants corresponding either to the full rationally
898 designed library ("full library"), the -1 PRF sites from the main set of events tested positive in our
899 assay combined ("frameshifting events") or to all variants based on one of these PRF events in
900 isolation. 10% of the variants in the individual sets (20% in case of the HIV variants) were put aside
901 and used only for evaluation of models built using the other 90% (80%). We chose Gradient Boosting
902 Decision Trees (XGBoost (Chen and Guestrin, 2016)) as the prediction algorithm because it can
903 capture non-linear interactions between features, allows combining different types of continuous and
904 categorical features and has proven to be a powerful approach in predicting the effect of regulatory
905 regions (Mikl et al., 2019; Vainberg Slutskin et al., 2018).

906 For predictions based on secondary structure we used the fold function from the Vienna RNA package
907 2.0 and extracted both the minimal free energy ("dg"; calculated for different lengths of downstream
908 (10-120 nt) regions) and the predicted pairedness for each position ("sec"; for each position in the
909 downstream region this set of features contains the information whether the position is predicted
910 (Vienna RNAfold) to be paired or not (0 or 1)).

911 For predictions based on tRNA adaptation index (tAI), we computed (based on Waldman et al., 2010)
912 the tAI of codons in the original and -1 frame as well as the (per position) difference between frame -
913 1 and frame 0 and used the information for 14 starting from the first codon in the slippery site (in the
914 case of the canonical -1 pattern XXXYYYZ, XXY in frame 0 and XXX in frame -1) as features.

915 For predictions based on amino acid identity, information on the type of amino acids around the
916 slippery site (from position -3 relative to the site of frameshifting until the end of the slippery site) was
917 used ("unpolar", "polar" or "charged").

Different hyperparameter settings for learning rate, n_estimators and max_depth were tested in a systematic and combinatorial fashion using 10-fold cross-validation for each set of variants and features independently. Typically around 100 tests were performed and the best set of hyperparameters used for subsequent steps.

At the end, the model was tested by training it on the entire training set (90% of all relevant unique library variants, 80% in the case of HIV clinical isolates) and scoring the correlation (Pearson) of predicted vs. measured values based on the held-out test set (10% of relevant unique library variants from the relevant set, 20% in the case of HIV clinical isolates), which had not been used at any stage during development of the model. This approach was used for each of the features sets in isolation as well as all of the features combined.

For classification, the same pipeline was used. Here, XGBClassifier was trained and the performance scored on the test set, reporting the area under the ROC curve as a metric.

Feature importance and effect on the model was determined using SHAP analysis (Lundberg et al., 2020).

General data analysis

For data analysis, we used python 2.7.11 with pandas 0.20.3, numpy 1.13.1, seaborn 0.6, scipy 0.17, scikit-learn 0.18.2 and shap 0.28.5. Confidence intervals were calculated by bootstrapping (1000 iterations).

Data availability

All data generated in this study are available in the NCBI gene expression omnibus (GEO) under accession GSE145684.

Code availability

The code used to process the data, generate the figures and train and test the predictive models will be made available as a GitHub repository (https://github.com/martinmikl/).

Acknowledgements

The authors thank Adina Weinberger, Orna Dahan, Alexey Gritsenko and Roni Rak for helpful discussions, Ronit Nir, Tali Avnit-Sagi and Maya Lotan-Pompan for technical advice and the Moshe Oren lab for advice and reagents. This work was supported by an EMBO long-term fellowship (to M.M.). E.S. is supported by the Crown Human Genome Center; the Else Kroener Fresenius Foundation; D. L. Schwarz; J. N. Halpern; L. Steinberg; J. Benattar; Aliza Moussaieff; Adelis Foundation; and grants funded by the European Research Council and the Israel Science Foundation.

Author contributions

Conceptualization: M.M., Y.P. and E.S.; Methodology, Software and Formal Analysis: M.M.; Investigation: M.M.; Writing: M.M., Y.P. and E.S.; Funding Acquisition: M.M., Y.P. and E.S.; Supervision: Y.P. and E.S.

The authors declare no competing interests.

References

Advani, V.M., and Dinman, J.D. (2016). Reprogramming the genetic code: The emerging role of ribosomal frameshifting in regulating cellular gene expression. BioEssays News Rev. Mol. Cell. Dev. Biol. 38, 21–26.

Baril, M., Dulude, D., Gendron, K., Lemay, G., and Brakier-Gingras, L. (2003). Efficiency of a programmed −1 ribosomal frameshift in the different subtypes of the human immunodeficiency virus type 1 group M. RNA 9, 1246–1253.

Belew, A.T., Hepler, N.L., Jacobs, J.L., and Dinman, J.D. (2008). PRFdb: a database of computationally predicted eukaryotic programmed -1 ribosomal frameshift signals. BMC Genomics 9, 339.

Belew, A.T., Advani, V.M., and Dinman, J.D. (2011). Endogenous ribosomal frameshift signals operate as mRNA destabilizing elements through at least two molecular pathways in yeast. Nucleic Acids Res. 39, 2799–2808.

Belew, A.T., Meskauskas, A., Musalgaonkar, S., Advani, V.M., Sulima, S.O., Kasprzak, W.K., Shapiro, B.A., and Dinman, J.D. (2014). Ribosomal frameshifting in the CCR5 mRNA is regulated by miRNAs and the NMD pathway. Nature 512, 265–269.

977 Bidou, L., Stahl, G., Grima, B., Liu, H., Cassan, M., and Rousset, J.P. (1997). In vivo HIV-1
978 frameshifting efficiency is directly related to the stability of the stem-loop stimulatory signal. RNA N.
979 Y. N 3, 1153–1158.

980 Biswas, P., Jiang, X., Pacchia, A.L., Dougherty, J.P., and Peltz, S.W. (2004). The Human
981 Immunodeficiency Virus Type 1 Ribosomal Frameshifting Site Is an Invariant Sequence Determinant
982 and an Important Target for Antiviral Therapy. J. Virol. 78, 2082–2087.

983 Brakier-Gingras, L., Charbonneau, J., and Butcher, S.E. (2012). Targeting frameshifting in the human
984 immunodeficiency virus. Expert Opin. Ther. Targets 16, 249–258.

985 Brierley, I., and Pennell, S. (2001). Structure and function of the stimulatory RNAs involved in
986 programmed eukaryotic-1 ribosomal frameshifting. Cold Spring Harb. Symp. Quant. Biol. 66, 233–
987 248.

988 Brierley, I., Digard, P., and Inglis, S.C. (1989). Characterization of an efficient coronavirus ribosomal
989 frameshifting signal: requirement for an RNA pseudoknot. Cell 57, 537–547.

990 Caliskan, N., Katunin, V.I., Belardinelli, R., Peske, F., and Rodnina, M.V. (2014). Programmed -1
991 frameshifting by kinetic partitioning during impeded translocation. Cell 157, 1619–1631.

992 Caliskan, N., Peske, F., and Rodnina, M.V. (2015). Changed in translation: mRNA recoding by -1
993 programmed ribosomal frameshifting. Trends Biochem. Sci. 40, 265–274.

994 Caliskan, N., Wohlgemuth, I., Korniy, N., Pearson, M., Peske, F., and Rodnina, M.V. (2017).
995 Conditional Switch between Frameshifting Regimes upon Translation of dnaX mRNA. Mol. Cell 66,
996 558-567.e4.

997 Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proc. 22nd ACM
998 SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD 16 785–794.

999 Chen, J., Petrov, A., Johansson, M., Tsai, A., O'Leary, S.E., and Puglisi, J.D. (2014). Dynamic
1000 pathways of −1 translational frameshifting. Nature 512, 328–332.

1001 Choi, J., O'Loughlin, S., Atkins, J.F., and Puglisi, J.D. (2020). The energy landscape of −1 ribosomal
1002 frameshifting. Sci. Adv. 6, eaax6969.

1003 Clark, M.B., Jänicke, M., Gottesbühren, U., Kleffmann, T., Legge, M., Poole, E.S., and Tate, W.P.
1004 (2007). Mammalian Gene PEG10 Expresses Two Reading Frames by High Efficiency –1
1005 Frameshifting in Embryonic-associated Tissues. J. Biol. Chem. 282, 37359–37369.

1006 ten Dam, E., Brierley, I., Inglis, S., and Pleij, C. (1994). Identification and analysis of the pseudoknot-
1007 containing gag-pro ribosomal frameshift signal of simian retrovirus-1. Nucleic Acids Res. 22, 2304–
1008 2310.

1009 Dinman, J.D. (2012). Control of gene expression by translational recoding. Adv. Protein Chem. Struct.
1010 Biol. 86, 129–149.

1011 Dinman, J.D., and Wickner, R.B. (1992). Ribosomal frameshifting efficiency and gag/gag-pol ratio are
1012 critical for yeast M1 double-stranded RNA virus propagation. J. Virol. 66, 3669–3676.

1013 Du, Z., Holland, J.A., Hansen, M.R., Giedroc, D.P., and Hoffman, D.W. (1997). Base-pairings within
1014 the RNA pseudoknot associated with the simian retrovirus-1 gag-pro frameshift site. J. Mol. Biol. 270,
1015 464–470.

1016 Felsenstein, K.M., and Goff, S.P. (1988). Expression of the gag-pol fusion protein of Moloney murine
1017 leukemia virus without gag protein does not induce virion formation or proteolytic processing. J. Virol.
1018 62, 2179–2182.

1019 Firth, A.E., and Atkins, J.F. (2009). A conserved predicted pseudoknot in the NS2A-encoding
1020 sequence of West Nile and Japanese encephalitis flaviviruses suggests NS1' may derive from
1021 ribosomal frameshifting. Virol. J. 6, 14.

1022 Firth, A.E., and Brierley, I. (2012). Non-canonical translation in RNA viruses. J. Gen. Virol. 93, 1385–
1023 1409.

1024 Gallant, J.A., and Lindsley, D. (1992). Leftward ribosome frameshifting at a hungry codon. J. Mol.
1025 Biol. 223, 31–40.

1026 Garcia-Miranda, P., Becker, J.T., Benner, B.E., Blume, A., Sherer, N.M., and Butcher, S.E. (2016).
1027 Stability of HIV Frameshift Site RNA Correlates with Frameshift Efficiency and Decreased Virus
1028 Infectivity. J. Virol. 90, 6906–6917.

1029 Girstmair, H., Saffert, P., Rode, S., Czech, A., Holland, G., Bannert, N., and Ignatova, Z. (2013).
1030 Depletion of Cognate Charged Transfer RNA Causes Translational Frameshifting within the
1031 Expanded CAG Stretch in Huntingtin. Cell Rep. 3, 148–159.

1032 Halma, M.T.J., Ritchie, D.B., Cappellano, T.R., Neupane, K., and Woodside, M.T. (2019). Complex
1033 dynamics under tension in a high-efficiency frameshift stimulatory structure. Proc. Natl. Acad. Sci. U.
1034 S. A.

1035 Hemelaar, J., Gouws, E., Ghys, P.D., and Osmanov, S. (2006). Global and regional distribution of
1036 HIV-1 genetic subtypes and recombinants in 2004. AIDS Lond. Engl. 20, W13-23.

1037 Hofacker, I.L., Priwitzer, B., and Stadler, P.F. (2004). Prediction of locally stable RNA secondary
1038 structures for genome-wide surveys. Bioinforma. Oxf. Engl. 20, 186–190.

1039 Hung, M., Patel, P., Davis, S., and Green, S.R. (1998). Importance of ribosomal frameshifting for
1040 human immunodeficiency virus type 1 particle assembly and replication. J. Virol. 72, 4819–4824.

1041 Ivanov, I.P., and Atkins, J.F. (2007). Ribosomal frameshifting in decoding antizyme mRNAs from
1042 yeast and protists to humans: close to 300 cases reveal remarkable diversity despite underlying
1043 conservation. Nucleic Acids Res. 35, 1842–1858.

1044 Ivanov, I.P., Gesteland, R.F., and Atkins, J.F. (1998). A second mammalian antizyme: conservation
1045 of programmed ribosomal frameshifting. Genomics 52, 119–129.

1046 Ivanov, I.P., Matsufuji, S., Murakami, Y., Gesteland, R.F., and Atkins, J.F. (2000). Conservation of
1047 polyamine regulation by translational frameshifting from yeast to mammals. EMBO J. 19, 1907–1917.

1048 Ivanov, I.P., Gesteland, R.F., and Atkins, J.F. (2006). Evolutionary specialization of recoding:
1049 Frameshifting in the expression of S. cerevisiae antizyme mRNA is via an atypical antizyme shift site
1050 but is still +1. RNA 12, 332–337.

1051 Jacks, T., and Varmus, H.E. (1985). Expression of the Rous sarcoma virus pol gene by ribosomal
1052 frameshifting. Science 230, 1237–1242.

1053 Jacks, T., Power, M.D., Masiarz, F.R., Luciw, P.A., Barr, P.J., and Varmus, H.E. (1988a).
1054 Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. Nature 331, 280–283.

1055 Jacks, T., Madhani, H.D., Masiarz, F.R., and Varmus, H.E. (1988b). Signals for ribosomal
1056 frameshifting in the Rous sarcoma virus gag-pol region. Cell 55, 447–458.

1057 Janssen, S., and Giegerich, R. (2015). The RNA shapes studio. Bioinformatics 31, 423–425.

1058 Ketteler, R. (2012). On programmed ribosomal frameshifting: the alternative proteomes. Front. Genet.
1059 3, 242.

1060 Kleinkauf, R., Houwaart, T., Backofen, R., and Mann, M. (2015a). antaRNA – Multi-objective inverse
1061 folding of pseudoknot RNA using ant-colony optimization. BMC Bioinformatics 16, 389.

1062 Kleinkauf, R., Mann, M., and Backofen, R. (2015b). antaRNA: ant colony-based RNA sequence
1063 design. Bioinformatics 31, 3114–3121.

1064 Kollmus, H., Honigman, A., Panet, A., and Hauser, H. (1994). The sequences of and distance between
1065 two cis-acting signals determine the efficiency of ribosomal frameshifting in human immunodeficiency
1066 virus type 1 and human T-cell leukemia virus type II in vivo. J. Virol. 68, 6087–6091.

1067 Korniy, N., Goyal, A., Hoffmann, M., Samatova, E., Peske, F., Pöhlmann, S., and Rodnina, M.V.
1068 (2019). Modulation of HIV-1 Gag/Gag-Pol frameshifting by tRNA abundance. Nucleic Acids Res. 47,
1069 5210–5222.

1070 Kurian, L., Palanimurugan, R., Gödderz, D., and Dohmen, R.J. (2011). Polyamine sensing by nascent
1071 ornithine decarboxylase antizyme stimulates decoding of its mRNA. Nature 477, 490–494.

1072 Lin, Z., Gilbert, R.J.C., and Brierley, I. (2012). Spacer-length dependence of programmed -1 or -2
1073 ribosomal frameshifting on a U6A heptamer supports a role for messenger RNA (mRNA) tension in
1074 frameshifting. Nucleic Acids Res. 40, 8674–8689.

1075 Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and
1076 Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. AMB 6, 26.

1077 Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J.,
1078 Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable AI
1079 for trees. Nat. Mach. Intell. 2, 56–67.

1080 Lyon, K., Aguilera, L.U., Morisaki, T., Munsky, B., and Stasevich, T.J. (2019). Live-Cell Single RNA
1081 Imaging Reveals Bursts of Translational Frameshifting. Mol. Cell 75, 172-183.e9.

1082 Manktelow, E., Shigemoto, K., and Brierley, I. (2005). Characterization of the frameshift signal of Edr,
1083 a mammalian example of programmed -1 ribosomal frameshifting. Nucleic Acids Res. 33, 1553–1563.

1084 Marcheschi, R.J., Staple, D.W., and Butcher, S.E. (2007). Programmed Ribosomal Frameshifting in
1085 SIV is Induced by a Highly Structured RNA Stem-Loop. J. Mol. Biol. 373, 652–663.

1086 Marczinke, B., Fisher, R., Vidakovic, M., Bloys, A.J., and Brierley, I. (1998). Secondary structure and
1087 mutational analysis of the ribosomal frameshift signal of rous sarcoma virus. J. Mol. Biol. 284, 205–
1088 225.

1089 Matsufuji, S., Matsufuji, T., Miyazaki, Y., Murakami, Y., Atkins, J.F., Gesteland, R.F., and Hayashi, S.
1090 (1995). Autoregulatory frameshifting in decoding mammalian ornithine decarboxylase antizyme. Cell
1091 80, 51–60.

1092 Melian, E.B., Hinzman, E., Nagasaki, T., Firth, A.E., Wills, N.M., Nouwens, A.S., Blitvich, B.J., Leung,
1093 J., Funk, A., Atkins, J.F., et al. (2010). NS1′ of Flaviviruses in the Japanese Encephalitis Virus
1094 Serogroup Is a Product of Ribosomal Frameshifting and Plays a Role in Viral Neuroinvasiveness. J.
1095 Virol. 84, 1641–1647.

1096 Melian, E.B., Hall-Mendelin, S., Du, F., Owens, N., Bosco-Lauth, A.M., Nagasaki, T., Rudd, S., Brault,
1097 A.C., Bowen, R.A., Hall, R.A., et al. (2014). Programmed ribosomal frameshift alters expression of
1098 west nile virus genes and facilitates virus replication in birds and mosquitoes. PLoS Pathog. 10,
1099 e1004447.

1100 Meydan, S., Klepacki, D., Karthikeyan, S., Margus, T., Thomas, P., Jones, J.E., Khan, Y., Briggs, J.,
1101 Dinman, J.D., Vázquez-Laslop, N., et al. (2017). Programmed Ribosomal Frameshifting Generates a
1102 Copper Transporter and a Copper Chaperone from the Same Gene. Mol. Cell 65, 207–219.

1103 Mikl, M., Hamburg, A., Pilpel, Y., and Segal, E. (2019). Dissecting splicing decisions and cell-to-cell
1104 variability with designed sequence libraries. Nat. Commun. 10, 4572.

1105 Moomau, C., Musalgaonkar, S., Khan, Y.A., Jones, J.E., and Dinman, J.D. (2016). Structural and
1106 Functional Characterization of Programmed Ribosomal Frameshift Signals in West Nile Virus Strains
1107 Reveals High Structural Plasticity Among cis-Acting RNA Elements. J. Biol. Chem. 291, 15788–
1108 15795.

1109 Moon, S., Byun, Y., and Han, K. (2007). FSDB: a frameshift signal database. Comput. Biol. Chem.
1110 31, 298–302.

1111 Mouzakis, K.D., Lang, A.L., Vander Meulen, K.A., Easterday, P.D., and Butcher, S.E. (2013). HIV-1
1112 frameshift efficiency is primarily determined by the stability of base pairs positioned at the mRNA
1113 entrance channel of the ribosome. Nucleic Acids Res. 41, 1901–1913.

1114 Nam, S.H., Copeland, T.D., Hatanaka, M., and Oroszlan, S. (1993). Characterization of ribosomal
1115 frameshifting for expression of pol gene products of human T-cell leukemia virus type I. J. Virol. 67,
1116 196–203.

1117 Nikolic, E.I.C., King, L.M., Vidakovic, M., Irigoyen, N., and Brierley, I. (2012). Modulation of ribosomal
1118 frameshifting frequency and its effect on the replication of Rous sarcoma virus. J. Virol. 86, 11581–
1119 11594.

1120 Ono, M., Yasunaga, T., Miyata, T., and Ushikubo, H. (1986). Nucleotide sequence of human
1121 endogenous retrovirus genome related to the mouse mammary tumor virus genome. J. Virol. 60, 589–
1122 598.

1123 Parkin, N.T., Chamorro, M., and Varmus, H.E. (1992). Human immunodeficiency virus type 1 gag-pol
1124 frameshifting is dependent on downstream mRNA secondary structure: demonstration by expression
1125 in vivo. J. Virol. 66, 5147–5151.

1126 Plant, E.P., Pérez-Alvarado, G.C., Jacobs, J.L., Mukhopadhyay, B., Hennig, M., and Dinman, J.D.
1127 (2005). A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. PLoS Biol.
1128 3, e172.

1129 Pollom, E., Dang, K.K., Potter, E.L., Gorelick, R.J., Burch, C.L., Weeks, K.M., and Swanstrom, R.
1130 (2013). Comparison of SIV and HIV-1 Genomic RNA Structures Reveals Impact of Sequence
1131 Evolution on Conserved and Non-Conserved Structural Motifs. PLOS Pathog. 9, e1003294.

1132 Ritchie, D.B., Foster, D.A.N., and Woodside, M.T. (2012). Programmed -1 frameshifting efficiency
1133 correlates with RNA pseudoknot conformational plasticity, not resistance to mechanical unfolding.
1134 Proc. Natl. Acad. Sci. U. S. A. 109, 16167–16172.

1135 Ritchie, D.B., Cappellano, T.R., Tittle, C., Rezajooei, N., Rouleau, L., Sikkema, W.K.A., and
1136 Woodside, M.T. (2017). Conformational dynamics of the frameshift stimulatory structure in HIV-1.
1137 RNA N. Y. N 23, 1376–1384.

1138 Rom, E., and Kahana, C. (1994). Polyamines regulate the expression of ornithine decarboxylase
1139 antizyme in vitro by inducing ribosomal frame-shifting. Proc. Natl. Acad. Sci. U. S. A. 91, 3959–3963.

1140 Staple, D.W., and Butcher, S.E. (2003). Solution structure of the HIV-1 frameshift inducing stem-loop
1141 RNA. Nucleic Acids Res. 31, 4326–4331.

1142 Su, M.-C., Chang, C.-T., Chu, C.-H., Tsai, C.-H., and Chang, K.-Y. (2005). An atypical RNA
1143 pseudoknot stimulator and an upstream attenuation signal for -1 ribosomal frameshifting of SARS
1144 coronavirus. Nucleic Acids Res. 33, 4265–4275.

1145 Theis, C., Reeder, J., and Giegerich, R. (2008). KnotInFrame: prediction of -1 ribosomal frameshift
1146 events. Nucleic Acids Res. 36, 6013–6020.

1147 Thiel, V., Ivanov, K.A., Putics, A., Hertzig, T., Schelle, B., Bayer, S., Weissbrich, B., Snijder, E.J.,
1148 Rabenau, H., Doerr, H.W., et al. (2003). Mechanisms and enzymes involved in SARS coronavirus
1149 genome expression. J. Gen. Virol. 84, 2305–2315.

1150 Tholstrup, J., Oddershede, L.B., and Sørensen, M.A. (2012). mRNA pseudoknot structures can act
1151 as ribosomal roadblocks. Nucleic Acids Res. 40, 303–313.

1152 Tosaka, Y., Tanaka, H., Yano, Y., Masai, K., Nozaki, M., Yomogida, K., Otani, S., Nojima, H., and
1153 Nishimune, Y. (2000). Identification and characterization of testis specific ornithine decarboxylase
1154 antizyme (OAZ-t) gene: expression in haploid germ cells and polyamine-induced frameshifting. Genes
1155 Cells Devoted Mol. Cell. Mech. 5, 265–276.

1156 Tu, C., Tzeng, T.H., and Bruenn, J.A. (1992). Ribosomal movement impeded at a pseudoknot
1157 required for frameshifting. Proc. Natl. Acad. Sci. U. S. A. 89, 8636–8640.

1158 Vainberg Slutskin, I., Weingarten-Gabbay, S., Nir, R., Weinberger, A., and Segal, E. (2018).
1159 Unraveling the determinants of microRNA mediated regulation using a massively parallel reporter
1160 assay. Nat. Commun. 9.

1161 Waldman, Y.Y., Tuller, T., Shlomi, T., Sharan, R., and Ruppin, E. (2010). Translation efficiency in
1162 humans: tissue specificity, global optimization and differences between developmental stages.
1163 Nucleic Acids Res. 38, 2964–2974.

1164 Wang, X., Xuan, Y., Han, Y., Ding, X., Ye, K., Yang, F., Gao, P., Goff, S.P., and Gao, G. (2019).
1165 Regulation of HIV-1 Gag-Pol Expression by Shiftless, an Inhibitor of Programmed -1 Ribosomal
1166 Frameshifting. Cell 176, 625-635.e14.

1167 Wang, Y., Wills, N.M., Du, Z., Rangan, A., Atkins, J.F., Gesteland, R.F., and Hoffman, D.W. (2002).
1168 Comparative studies of frameshifting and nonframeshifting RNA pseudoknots: a mutational and NMR
1169 investigation of pseudoknots derived from the bacteriophage T2 gene 32 mRNA and the retroviral
1170 gag-pro frameshift site. RNA N. Y. N 8, 981–996.

1171 Weingarten-Gabbay, S., Elias-Kirma, S., Nir, R., Gritsenko, A.A., Stern-Ginossar, N., Yakhini, Z.,
1172 Weinberger, A., and Segal, E. (2016). Systematic discovery of cap-independent translation
1173 sequences in human and viral genomes. Science 351, aad4939.

1174 Weingarten-Gabbay, S., Nir, R., Lubliner, S., Sharon, E., Kalma, Y., Weinberger, A., and Segal, E.
1175 (2019). Systematic interrogation of human promoters. Genome Res. 29, 171–183.

1176 Weiss, R., Lindsley, D., Falahee, B., and Gallant, J. (1988). On the mechanism of ribosomal
1177 frameshifting at hungry codons. J. Mol. Biol. 203, 403–410.

1178 van Weringh, A., Ragonnet-Cronin, M., Pranckeviciene, E., Pavon-Eternod, M., Kleiman, L., and Xia,
1179 X. (2011). HIV-1 modulates the tRNA pool to improve translation efficiency. Mol. Biol. Evol. 28, 1827–
1180 1834.

1181 Wills, N.M., Moore, B., Hammer, A., Gesteland, R.F., and Atkins, J.F. (2006). A functional -1 ribosomal
1182 frameshift signal in the human paraneoplastic Ma3 gene. J. Biol. Chem. 281, 7082–7088.

1183 Wu, B., Zhang, H., Sun, R., Peng, S., Cooperman, B.S., Goldman, Y.E., and Chen, C. (2018).
1184 Translocation kinetics and structural dynamics of ribosomes are modulated by the conformational
1185 plasticity of downstream pseudoknots. Nucleic Acids Res. 46, 9736–9748.

1186 Yan, S., Wen, J.-D., Bustamante, C., and Tinoco, I. (2015). Ribosome excursions during mRNA
1187 translocation mediate broad branching of frameshift pathways. Cell 160, 870–881.

1188 Yordanova, M.M., Wu, C., Andreev, D.E., Sachs, M.S., and Atkins, J.F. (2015). A Nascent Peptide
1189 Signal Responsive to Endogenous Levels of Polyamines Acts to Stimulate Regulatory Frameshifting
1190 on Antizyme mRNA. J. Biol. Chem. 290, 17863–17878.

1191