Extensive programmed ribosomal frameshifting in human as revealed by a massively parallel reporter assay

Martin Mikl[123*], Amit Alon[3], Ernest Mordret[3], 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-CoV and is critical for their life cycle and virulence. It is also utilized throughout the tree of life to implement a feedback control mechanism to regulate polyamine levels. However, despite its universality and clinical relevance, a limited number of studies investigated this process on only a few selected examples, largely due to a lack of experimental means. Here, we developed a high-throughput, fluorescence-based approach to assay 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 features. We also utilized our approach to search for novel frameshifting events and identified dozens of previously unknown frameshifting sites in human, showing that programmed ribosomal frameshifting is more common than previously anticipated. 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 a machine learning algorithm that accurately predicts frameshifting rates of novel variants (up to r=0.70), including subtle differences between HIV isolates (r=0.44), providing a basis for 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. This is a strikingly robust and precise process with minimal variability (Dulude et al., 2006; Hung et al., 1998), which might explain why viruses rely on it rather than other means of posttranscriptional and translational control for crucial regulatory switches. 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; Tosaka et al., 2000). 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 (Kurian et al., 2011; Matsufuji et al., 1995). 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.

Frameshifting 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 XXXYYYZ, with the shift happening from codon YYZ to YYY. Some known slippery sites show more or less extensive divergence from this pattern, and especially at +1 frameshifting sites like OAZ ribosomal translocation happens at a very distinct motif (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 happens and the prevalence of the process in eukaryotes remain largely unknown.

Here, we developed a massively parallel reporter assay that allows for high throughput quantification of ribosomal frameshifting in human cells. We designed and tested 17,809 oligonucleotides containing

rationally designed variants of known frameshifting signals as well as a large collection of native sequences suspected to have the ability to induce ribosomal frameshifting. We systematically deciphered determinants of PRF efficiency across frameshifting events, identified >100 novel frameshifting events in human and viral genomes and assayed natural variation in HIV gag-pol frameshifting, providing the first systematic large-scale investigation of ribosomal frameshifting.

Results and Discussion

A massively parallel reporter assay accurately measures PRF rates and provides evidence for bidirectional frameshifting

To assay PRF in a comprehensive manner we designed a synthetic oligonucleotide library containing (a) 12809 variants with systematic sequence manipulations of previously reported frameshift sites, (b) 4019 native viral and human sequences suspected to have the ability to induce frameshifting based on present predictions (Belew et al., 2008), dual coding potential (Chung et al., 2007), ribosome profiling data (Ingolia et al., 2009) and a novel approach detecting unexpected patterns of sequence conservation (Alon et al., manuscript in preparation), and (c) 581 sequences of gag-pol frameshifting sites in HIV clinical isolates (collated from http://www.hiv.lanl.gov/) (Fig 1A).

The oligonucleotides comprising library-specific common primers, a unique barcode and a 162 nt long variable region were synthesized on an Agilent microarray, amplified and cloned in between *mcherry* and *gfp* coding sequence such that the *gfp* coding frame was shifted by +1 or -1 relative to the original, mCherry-encoding frame (Fig 1B). GFP would only be made into protein if the corresponding frameshift occurred, and GFP fluorescence intensity thus serves as a measure for frameshifting efficiency. We introduced this construct in the AAVS1 locus in the human K562 cell line using zinc finger nucleases, such that every cell has one frameshifting reporter construct from the library and all the variants have the same genomic environment (Methods). For both, the -1 and +1 reporter libraries, we sorted the mCherry-positive population corresponding to a single integration of the reporter transgene using flow cytometry into 16 bins according to their GFP fluorescence intensity and sequenced genomic DNA from all the bins to unravel the distribution of each variant. We previously demonstrated that similar approaches are highly accurate and reproducible (Mikl et al., 2018; Vainberg Slutskin et al., 2018; Weingarten-Gabbay et al., 2016), and the bin profiles for barcode control groups with identical variable region (Fig S1A) corroborate the low technical noise we are able to achieve. Moreover, by measuring GFP fluorescence of our reporter construct containing the human OAZ1 frameshifting site in the presence of different levels of spermidine we confirmed that we were able to detect the previously reported sensitivity to polyamine levels (Kurian et al., 2011) using our FACS-based reporter assay (Fig S1B).

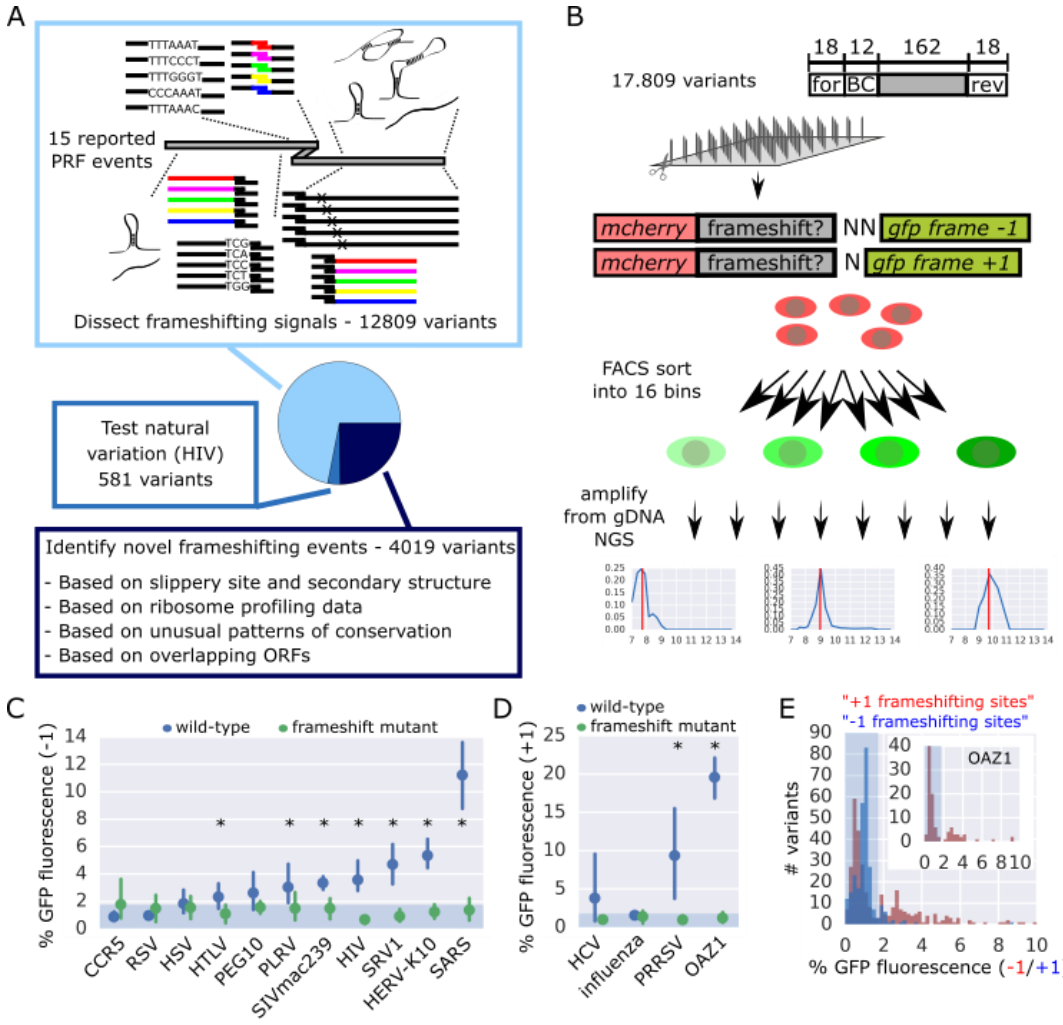

Figure 1. A massively parallel reporter assay quantitatively measures PRF rates and reveals bidirectional shifting.

A. Schematics of the library design. B. Outline of the experimental pipeline; for: forward primer, rev: reverse primer, BC: barcode. CD. Mean and 95% CI for barcode control groups corresponding to the indicated wild-type -1 (C) and +1 (D) PRF sequence (blue) or carrying point mutations in the slippery site (green). E. Histogram of mean -1 (red) or +1 (blue) % GFP fluorescence for variants based on frameshift events annotated as shifting in the opposite direction, i.e. +1 (red) or -1 (blue).

The distribution of mean GFP expression of all library variants shows clear peaks corresponding to background green fluorescence and maximum GFP levels from variants with GFP in frame with mCherry (Fig S1C). We set the lowest mean GFP fluorescence we observed to 0% and the highest to 100%. Accordingly, we assigned a percentage to every variant that passed filtering for read number, bin profile and 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 necessarily denote the precise rate of frameshifting events, but gives us a meaningful measure of frameshifting efficiencies across PRF events. Based on the peak of negative, non-frameshifting variants we assigned a noise threshold (corresponding to 1.7%

of the maximal GFP fluorescence) imposed by autofluorescence of the cells. We also excluded variants exhibiting a dominant peak with mean GFP fluorescence above 2^12 (25%) to rule out biases stemming from DNA frameshifts occurring as synthesis or cloning errors and leading to GFP being in frame with mCherry and thereby alleviating the need for ribosomal frameshifting to achieve a fluorescent signal (examples are shown in Fig S1D-I; see also Methods).

We tested previously reported frameshifting sites (with or without experimental validation, Table S1, based on Moon et al., 2007 and our own survey of the literature) with multiple different barcodes in our assay and could reproducibly detect GFP fluorescence in the expected frame in many of the cases tested (Fig 1CD, Fig S2A). Additional requirements for frameshifting (like the presence of a specific miRNA as in the case of CCR5 (Belew et al., 2014)) can explain the lack of signal for some of the cases. Most single point mutations within the slippery sequence abrogated frameshifting and resulted in green fluorescence at background levels, both for -1 and +1 frameshifting sites (Fig 1CD, Fig S2B), demonstrating that we indeed measure frameshifting at the expected site. Notably, a cytosine in position 4 in the canonical slippery site XXXYYYZ – although disrupting the decoding compatibility between the source and the target codon – did not interfere with PRF in three out of five cases (Fig S2C), against the naïve assumption that identity in the first position in the codon at which the shift to the -1 frame happens would be critical for the efficiency of the process. Replacing the slippery sequence in our set of -1 PRF events with all possible variations of the pattern XXXYYYZ revealed preferences for specific combinations common in known PRF sites, like UUUUUUZ (HIV, SIVmac239) and XXXAAAC (HERV-K10, HTLV, PEG10, SARS), but no general inhibition of PRF by the presence of a specific base at any of the three positions (Fig S2D).

Frameshifting in the 5’ or 3’ directions are generally thought to be mutually exclusive and previously reported -1 and +1 frameshifting sites show distinct characteristics in their slippery site (XXXYYYZ, or variations thereof, for -1; a stop codon in the case of OAZ (+1 PRF)). Using our comprehensive assay we observed that -1 slippery sites generally lack the ability to frameshift in the 3’ direction (and therefore do not give a signal in the +1 frame), whereas +1 slippery sites do have the ability to induce -1 frameshifting (Fig 1E). Strikingly, this includes the OAZ1 frameshifting site, which is fundamentally different from all known -1 PRF slippery sites. Native OAZ1 has several stop codons in the -1 frame downstream of the frameshifting site and -1 frameshifted versions can therefore not be detected in our assay, but specific combinations of upstream and downstream native sequences from other frameshifting sites (and lacking stop codons in the -1 frame) have the ability to drive -1 frameshifting at the native OAZ1 slippery site (Fig S2E). These data suggest that the ability to induce -1 frameshifting might be a general and more basal property of any potential frameshifting site.

To confirm that our assay indeed measures frameshifting, we pulled down GFP (and mCherry) containing molecules from the complete pool of 17809 variants and aimed to identify transframe

peptides (i.e. peptides from PRF products that span the frameshifting site) by mass spectrometry (Fig S3A). Despite the complexity of the library and the caveats associated with trying to detect specific peptides in mass spectrometry data, we identified 11 peptides not present in the human proteome and representing the product of a frameshifting event. Among others, we detected a transframe peptide from a variant of the herpes simplex frameshifting site with two nucleotides introduced after the slippery site, which – unlike the corresponding wild-type sequence – also shows a frameshifting signal in our FACSseq assay (Fig S3B), confirming that we indeed identify specific sequence variants able to induce frameshifting.

Sequence, structural and amino acid properties affect PRF efficiency

While previous investigations of frameshifting focused on individual examples, we aimed to identify commonalities and differences between frameshifting sites. While mutations in the slippery site and the downstream region have a negative effect on frameshifting efficiency across contexts, changes in the upstream region tend to lead to higher -1 frameshifting rates (between 9.5% and 45.9% mean increase compared to wild-type frameshifting rates for different PRF events, Fig 2A; p-values per position (Mann-Whitney U test) shown in Fig S4A). This suggests that inhibitory signals upstream of the frameshifting site as found for SARS (Su et al., 2005) could be a widespread property of PRF sites, probably creating a balance to ensure that frameshift promoting signals like a rigid downstream secondary structure don't lead to a complete inhibition of translation and potentially degradation of the mRNA. In contrast to the other -1 PRF events assayed, the native region upstream (but not downstream) of the HTLV PRF site enhances frameshifting. Consequently, changes in the upstream region in general (Fig 2A, Fig S4A, mean decrease of 23% compared to wild-type frameshifting rates) and recoding or manipulating the secondary structure of the upstream region (Fig 2B, mean decrease of 58.9% compared to wild-type frameshifting rates, $p<5*10^{-4}$) had strong negative effects on the downstream PRF event. While the upstream region of OAZ1 is thought to harbor positive signals (cf. Ivanov et al., 2006), which we find to be located in the 20 nucleotides before the slippery site (Fig 2C, decrease of PRF activity to 31.3% of wild-type PRF rate, $p<0.05$, Wilcoxon signed-rank test, effect on region further upstream not significant), HIV frameshifting rates show an increase upon mutation of the corresponding upstream region (Fig 2BC, mean increase of PRF activity to 128.2% of wild-type rates, $p<0.007$, Wilcoxon signed-rank test, effect on region further upstream not significant).

Preferences for downstream positions to be paired or unpaired reveal the properties of structural elements downstream of the PRF site. Groups of frameshifting events show remarkable concordance between these preferences (e.g. HIV, SARS and SIVmac239, Fig 2D), but length and position of the optimal downstream secondary structures differ between groups (Fig 2Dt, upper vs. lower panel, Fig S4B). Scanning mutagenesis revealed large differences in the extent of the relevant downstream region (Fig 2E), ranging from 10% of single point downstream reducing -1 PRF efficiency by more

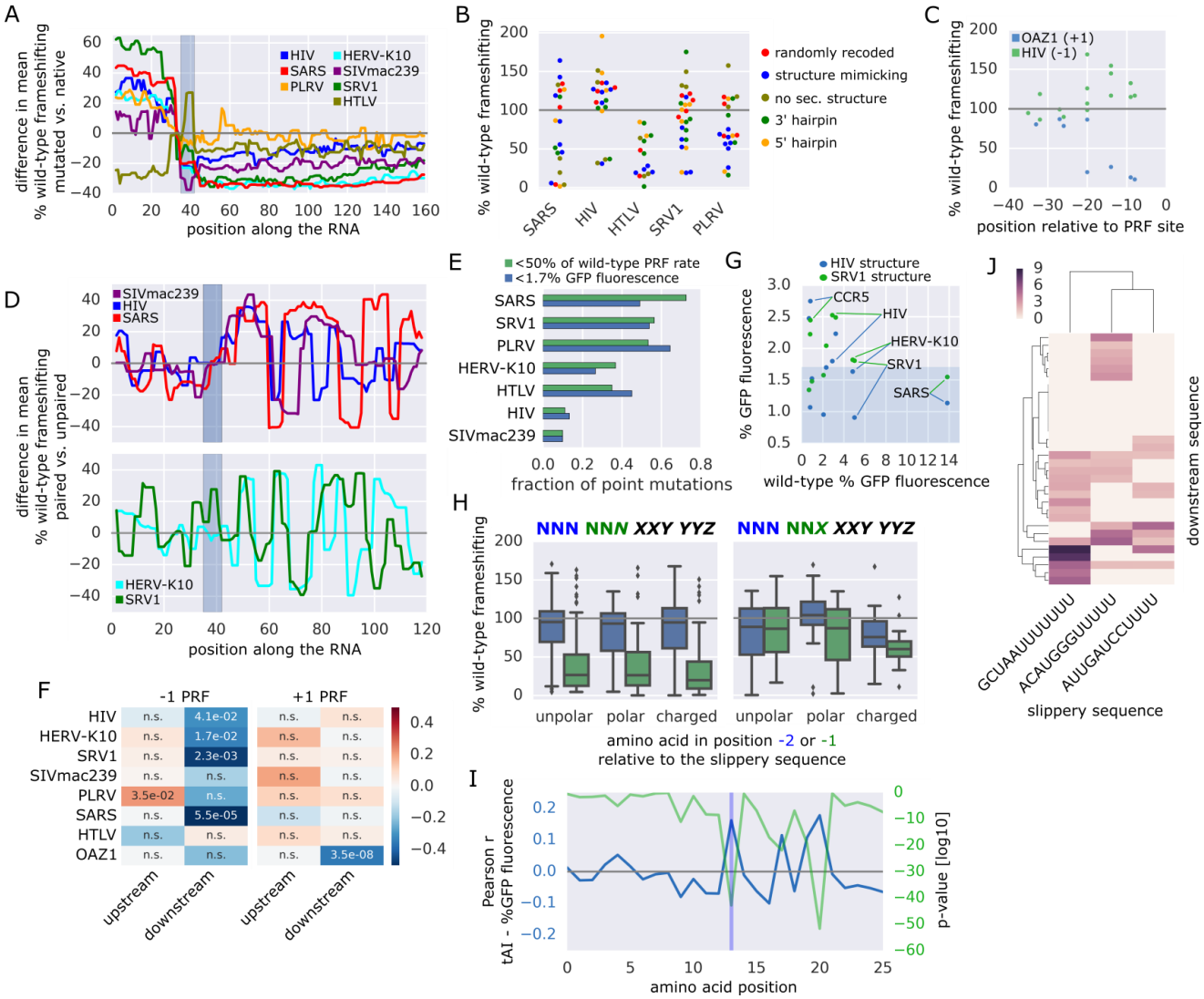

Figure 2. Sequence, structural and amino acid properties affecting PRF efficiency (legend on the next page).

than half in the case of SIVmac239 to 73% in the case of SARS (54% for +1 PRF at the OAZ1 site), and demonstrated that – on average – higher frameshifting efficiency also entails greater sensitivity to even minimal sequence changes (e.g. SARS vs HIV, Fig 2E, Fig S4C). The HIV gag-pol PRF site on the other hand shows remarkable resilience to point mutations (Fig S4C), in line with the high degree of genetic variability in HIV. Changes in secondary structure induced by sequence alterations downstream of the slippery site showed strong correlation with some, but not all frameshifting events (Fig 2F). Replacing the endogenous downstream region with elements that have different primary sequence, but are predicted to fold into the same secondary structure abolished frameshifting in most cases (Fig S4DE). Introducing variants of the SRV1 and to a lesser extent the HIV downstream structure (without preserving the original sequence) had the ability to trigger frameshifting in some cases, including ones where the wild-type sequence did not show frameshifting potential in our assay (Fig 2G, Fig S4F). Strikingly, events associated with higher wild-type frameshifting rates could not be

Figure 2. Sequence, structural and amino acid properties affecting PRF efficiency.

A. The difference in mean % wild-type frameshifting between variants in which the indicated position is mutated vs non-mutated is plotted for the entire length of the variable region; gray box: slippery sequence. B. Boxplot of % wild-type frameshifting rates for variants with the sequence upstream of the slippery site being randomly recoded or replaced with sequences predicted to have the indicated secondary structure. C. % 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. 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 sequence. E. Fraction of point mutations in the 40 nt downstream of the PRF site resulting in <50% of wild-type PRF rates (green) and background fluorescence (<1.7% GFP fluorescence, blue). F. Heat map showing the Pearson correlation coefficient (and annotated with the corresponding p value) between +1 or -1 PRF rates and the minimum free energy of the 20 nt upstream and 40 nt downstream region, respectively. G. Mean % GFP fluorescence of variants in which the downstream region was replaced with sequences (n=3-10 for each data point) resembling either the HIV or the SRV1 secondary structure, plotted against the corresponding wild-type values. H. 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 (right) or not (left). I. Pearson correlation coefficient (blue) and associated p-values (green) between tAI at the indicated position and percent GFP fluorescence. J. Clustered heat map showing all possible combinations of 3 synthetic slippery sites and 34 synthetic downstream variants (minimal value 1.7%).

“rescued”, indicating that more efficient frameshifting events like SARS seem to be highly optimized, but less tolerant to sequence changes. In general, although present secondary structure prediction algorithms might not be sufficiently accurate for this task, especially in the case of pseudoknots, our results support a view according to which not only the structure, but also the sequence downstream of a slippery site is critical for triggering frameshifting.

We expanded our search for properties affecting PRF efficiency and examined the effect of the codons preceding and following the slippery site. We found that the presence of a charged amino acid immediately upstream of the slippery site reduced frameshifting efficiency approximately by half on average, even when the sequence of the slippery site was unchanged (Fig 2H, $p<0.0015$, Wilcoxon signed-rank test). In contrast, the first codon downstream of the secondary site showed purely DNA sequence, but not amino acid-specific effects (Fig S4G, exemplified by the differences between the Proline codons CCA, CCC and CCG). We further examined the influence of decoding efficiency as measured by the tRNA adaptation index (tAI) on PRF and found that the tAI of the codon at which the translocation happens is positively correlated with PRF efficiency (Fig 2I), indicating that decoding efficiency at the shifting codon is not contributing to stalling of the ribosome.

To not limit ourselves to endogenous frameshifting events, we designed partly and completely synthetic regions with different sequence and structural properties. Some fully designed downstream regions were able to induce frameshifting. Typically these were not the ones with the most stable

secondary structure (Fig S4H), corroborating our earlier findings that a stable secondary structure downstream of a slippery site alone is not sufficient to induce frameshifting. Accordingly, when we tested all different downstream regions with synthetic slippery sequences resembling common types of -1 PRF sites we found pronounced differences in combinatorial preferences (Fig 2J), showing that we only start to understand the full extent of the complexity of frameshifting regulation.

PRF is common in the human genome and happens at characteristic rates

Having accurate quantitative measurements for large collections of frameshifting sites, we aimed to predict frameshifting efficiency using machine learning tools. Based on our findings, we used tAI of codons around the frameshifting site, amino acid class, minimum free energy and pairedness of positions downstream of the frameshifting site, alone and in combination, as features and trained a Gradient Boosting Regressor. We achieved high accuracy (up to Pearson r=0.7) when training our model on variants of specific frameshifting events and predicting unseen variants from the same event (Fig 3A). Given the lack of ubiquitously valid features determining frameshifting rates, quantitative prediction for completely unrelated sequences can be expected to be less successful than mining available data and annotations to identify novel candidates for PRF. We assembled a list of 4019 candidates based on independent data (slippery site of type XXXYYYZ, followed by a stable

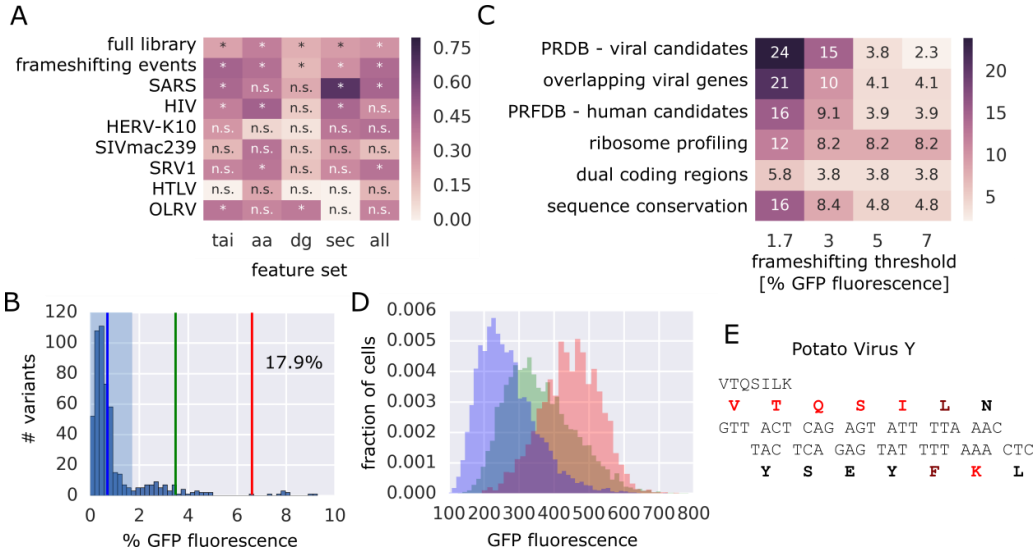

Figure 3. Prediction of known PRF events and identification of novel human and viral candidates.

A. Prediction score (r) on held-out test data for the indicated PRF event(s) or the entire library using different feature sets (tAI, aa: amino acid class ("unipolar", "polar", "charged"), dg: MFE of upstream and downstream regions, sec: pairedness of downstream positions). B. Distribution of % GFP fluorescence of all measured human and viral candidates; the vertical lines denote the values corresponding to the variants in panel D. C. The percentage of tested variants passing the indicated thresholds for the different subsets. D. Distribution of single cell GFP fluorescence read by FACS of three isolated clones, for which the mean value is indicated in panel B. E. Example for a mass-spectrometry identified peptide spanning a frameshift site.

secondary structure (based on PRFDB, Belew et al., 2008); overlapping ORFs or dual coding regions; evidence from ribosome profiling data for translation in a non-canonical frame (Ingolia et al., 2009), and unexpected patterns of sequence conservation (Alon et al., manuscript in preparation)) and tested their frameshifting potential. Around 18% of candidates gave a frameshifting signal above the noise threshold (Fig 3B), with a peak around the PRF rates of many known sites (2.5-3%), indicating that there is a typical basal rate of frameshifting that is commonly found in viral and human genomes. Candidates from ribosome profiling data showed a tendency to have higher frameshifting rates (Fig 3C), consistent with the assumption that only more abundant translation events in alternative frames would be picked up. Measuring individual library clones in isolation resulted in good correlation with the FACSseq measurements and – together with mass spectrometry-identified peptides mapping to frameshifting sites or the -1 frame (Fig 3DE, Fig S5A-D) – corroborated that we indeed identified a novel set of frameshifting events (Tables S2, S3 and S4).

Many viral candidates exhibited several possible slippery sites and therefore potential PRF events in tandem (21 at the minimal distance of 3 bp and 45 less than 100 bp away, out of 135 predicted by PRFDB (Belew et al., 2008) and tested here). The moderate PRF rates observed for these cases (e.g in the case of Simian immunodeficiency virus SIV-mnd 2 and Turkey astrovirus, Table S2) present this as a backup mechanism to ensure frameshifting rather than a strategy to increase overall frameshifting rates.

Human genes with frameshifting sites identified here (Table S4) are significantly enriched for the GO terms “response to folic acid” ($p<4*10^{-4}$, as determined using Gorilla (Eden et al., 2009) and “translational repressor activity” ($p<7*10^{-4}$). Although this enrichment does not pass multiple testing correction due to the low number of genes in the set, it highlights one gene associated with both GO terms – thymidylate synthase (TYMS, Fig S5E) – as a particularly promising candidate for a functional role of PRF in the cell. TYMS functions in nucleotide biosynthesis and is critical for maintaining the dTMP pool in the cell. It is a common oncogene (Rahman et al., 2004) and its functioning is influenced by an autoregulatory negative feedback loop as well as folate and dUMP levels (Chu et al., 1991), reminiscent of polyamine-stimulated frameshifting in OAZ. Given the extraordinary evolutionary conservation of the OAZ PRF it would not be surprising if similar strategies had evolved in the sensing of other compounds.

PRF rates of HIV clinical isolates exhibit subtype-specific differences and associations with viral load in patients

The HIV gag-pol frameshifting site is arguably one of the most intensely studied examples of PRF and – due to its critical importance for the viral replication cycle – has been repeatedly suggested as an antiviral drug target (Brakier-Gingras et al., 2012; Hung et al., 1998). To assess the natural variation in frameshifting rates in HIV1 we assembled a set of 581 sequences from clinical isolates

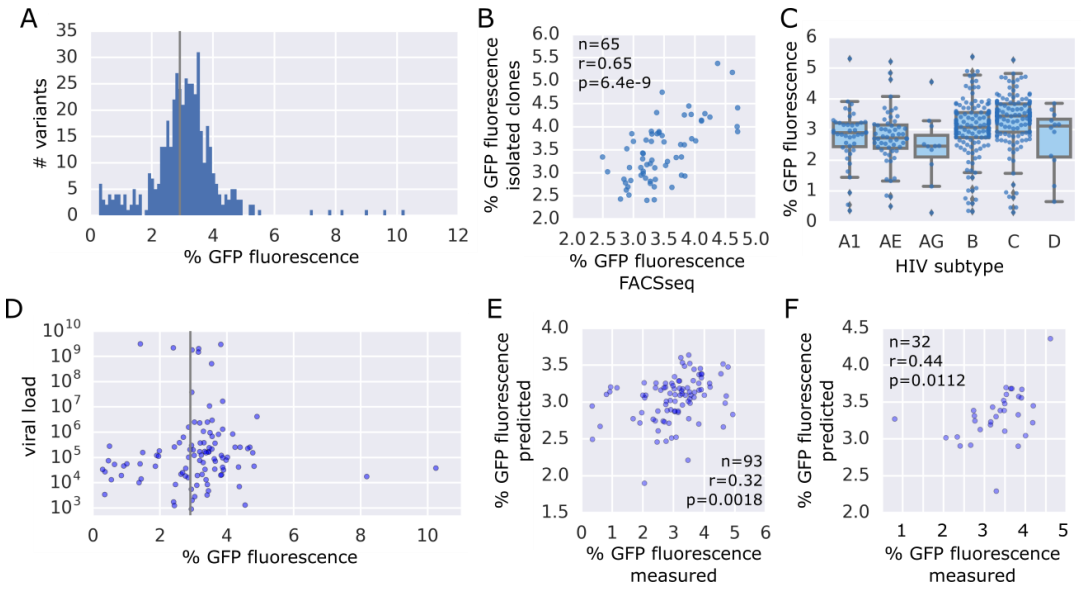

Figure 4. Testing of PRF sites from HIV clinical isolates reveals subtype specific differences and associations with viral load in patients.

A. Distribution of % GFP fluorescence of all measured HIV variants. B. % GFP fluorescence as determined by FACSseq is plotted against the value determined by measuring the corresponding isolated clone by FACS. C. Boxplot showing % GFP fluorescence of HIV gag-pol PRF variants coming from the indicated subtypes. D. Viral load (HIV titer) is plotted against % GFP fluorescence determined for the PRF site variant from the corresponding clinical isolate; vertical line: HIV HXB2 wild-type % GFP fluorescence. EF. Predicted vs measured % GFP fluorescence for 20% of the HIV variants (held-out test set) without (E) or with (F) additional filtering for variants with only one peak in their raw bin profiles.

between 1976 and 2014 (http://www.hiv.lanl.gov/), which differ in the sequence surrounding the frameshifting site, but not in the slippery site itself (Fig S6A). Using mass spectrometry we detected peptides covering most of frame 0 and frame -1 after the frameshifting site, with peptides coming from multiple variants (Fig S6A). Frameshifting rates of the variants are distributed around the rate observed for the lab wild-type strain HXB2 (Fig 4A). This constitutes actual differences in frameshifting rates and not only experimental variability, as isolated clones show remarkably good correlation given the small differences in PRF rates we are measuring (Fig 4B). Secondary structure showed the best correlation with frameshifting rates when considering the first 40 nucleotides after the frameshifting site (Fig S6B), matching the region of high sequence conservation (Fig S6A). We grouped the HIV variants based on subtype and found significant differences between the groups (Fig 4C, $p<3*10^{-5}$, one-way ANOVA), most notably higher frameshifting rates in subtype C ($p<0.006$ for the difference between C and B). HIV subtypes show distinct geographical distributions (Hemelaar et al., 2006), and consequently we also observed differences between countries of origin (Fig S6C, $p<0.002$), but no change in frameshifting rates over time (Fig S6D, $p=0.8$).

Optimal gag-pol frameshifting rates have been proposed to be critical for virulence (Dulude et al., 2006; Hung et al., 1998). In order to link the frameshifting rate measured in our reporter assay with

the replicative success of the corresponding HIV isolate we compared the viral load in patients (where available) to frameshifting rates of the causal HIV isolate (Fig 4D). 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 (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.

We identified the most predictive combination of features based on our results from the designed data set (Fig 3A, pairedness of nucleotides after the frameshifting site and amino acid class) and trained a Gradient Boosting Regressor on 80% of our HIV clinical isolates. Our prediction for 93 novel HIV variants showed good correlation with the experimentally measured values (r=0.32, Fig 4E), which could be further improved by applying more restrictive filtering (r=0.44, Fig 4F; see Methods). This shows that a model based on data from our assay is sensitive enough to detect even subtle differences between HIV variants and allows prediction of frameshifting rates with an accuracy of clinical relevance.

Conclusion

In summary, we combined the power of fluorescent frameshifting reporters with rational design of DNA sequences and high-throughput testing to systematically decipher the rules governing PRF and to identify novel candidates for a function of PRF in human. By controlled sequence and structure manipulations in multiple contexts we were able to dissect and directly compare the regulatory architecture of 15 frameshifting events, revealing a great diversity in regulatory strategies involving upstream and downstream sequence and structural elements and amino acid properties. Moreover, we found that sites reported to have +1 frameshifting potential, including OAZ, can also trigger a -1 (or +2) frameshift and identified differential preferences for combinations of slippery sites and downstream secondary structures.

Only very few PRF events encoded in the human genome have been identified to date. Here, we provide evidence for 54 additional frameshifting events, and more generally for the notion that ribosomal frameshifting could be a much more widespread phenomenon, corroborating hypotheses according to which up to 10% of human genes might be regulated by programmed ribosomal frameshifting (Advani and Dinman, 2016). In addition, our extensive library of frameshifting reporters provides a platform for screening putative modifiers of ribosomal frameshifting, offering a powerful tool to identify ways to generally or selectively control frameshifting and opening new possibilities for interfering with viral replication and for controlling cellular processes depending on translational frameshifting.

Acknowledgements

The authors thank Adina Weinberger, Orna Dahan, Alexey Gritsenko and Roni Rak for helpful discussions and Ronit Nir, Tali Avnit-Sagi and Maya Lotan-Pompan for technical advice. 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., A.A. and E.M.; Investigation: M.M.; Writing: M.M., Y.P. and E.S.; Funding Acquisition: M.M. 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.

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., 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.

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

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

Chu, E., Koeller, D.M., Casey, J.L., Drake, J.C., Chabner, B.A., Elwood, P.C., Zinn, S., and Allegra, C.J. (1991). Autoregulation of human thymidylate synthase messenger RNA translation by thymidylate synthase. Proc. Natl. Acad. Sci. *88*, 8977–8981.

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

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

Dulude, D., Berchiche, Y.A., Gendron, K., Brakier-Gingras, L., and Heveker, N. (2006). Decreasing the frameshift efficiency translates into an equivalent reduction of the replication of the human immunodeficiency virus type 1. Virology *345*, 127–136.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics *10*, 48.

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

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

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science *324*, 218–223.

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

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

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

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

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

Mikl, M., Hamburg, A., Pilpel, Y., and Segal, E. (2018). Dissecting splicing decisions and cell-to-cell variability with designed sequence libraries. BioRxiv 392605.

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

Rahman, L., Voeller, D., Rahman, M., Lipkowitz, S., Allegra, C., Barrett, J.C., Kaye, F.J., and Zajac-Kaye, M. (2004). Thymidylate synthase as an oncogene: a novel role for an essential DNA synthesis enzyme. Cancer Cell *5*, 341–351.

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

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

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

Tosaka, Y., Tanaka, H., Yano, Y., Masai, K., Nozaki, M., Yomogida, K., Otani, S., Nojima, H., and Nishimune, Y. (2000). Identification and characterization of testis specific ornithine decarboxylase

antizyme (OAZ-t) gene: expression in haploid germ cells and polyamine-induced frameshifting. Genes Cells Devoted Mol. Cell. Mech. *5*, 265–276.

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

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