Supplementary Note 1: Adapted RESCUE strategy

To handle reads mapped to more than one genomic location, we implemented an adopted RESCUE algorithm [Mortazavi et al., 2008]. Clusters of reads were built by first identifying groups of overlapping reads and then merging all clusters within the same gene. Clusters with less than 5 collapsed reads (treat all reads mapping to the same genomic coordinates as one) and less than 10 reads were discarded. Ambiguous alignments were resolved by applying the following three steps:

1. If a read maps to more than one cluster, the number u_i of uniquely mapping reads within all these clusters is considered. All alignments to clusters with $u_i < 0.1 \max(u_i)$ are discarded.

2. For all remaining alignment, the coverages of reads within its genomic coordinates are computed (with fractional counts for ambiguous alignments, i.e. $1/n$ with n alignments). For each alignment the minimal coverage c_i is then used: All alignments with $c_i < 0.1 \max(c_i)$ are discarded.

3. For all remaining alignments, the context is determined as the coverage of reads 100 bp upstream and downstream of the alignment (with fractional counts for ambiguous alignments; when there is a splice site within this context, all annotated splice junctions are considered and the splice partner with the most reads is used). For each alignment the median context t_i is then used: If $max(t_i) < 0.01C$, with C is the uncollapsed read count of the alignment (i.e. the number of reads aligned to the genomic coordinates), all alignments are discarded. Otherwise, all alignments with $t_i < 0.1 \max(ti)$ are discarded.

If more than one alignment remains, fractional counts are used for all subsequent steps unless otherwise noted. As the HSV-1 genome has two inverted repeats with identical sequences, terminal repeats were replaced by N, facilitating the resolution of the corresponding ambiguous mappings of reads.

Supplementary Note 2: Estimating cleavage and untemplated addition probabilities

The parameters of the model are the probability for an untemplated addition a, the probabilities for upstream cleavage positions at a distance of p from the P site u_1 through u_M with $\sum_{p=1}^{M} u_p = 1$ and the probabilities for downstream cleavage positions at a distance of p from the P site d_1 through d_M with $\sum_{p=1}^{M} d_p = 1$. Thus, the parameters to estimate are $\theta = (a, u_1, ..., u_M, d_1, ..., d_M)$.
Assuming that upstream and downstream cleavage is independent from each other, and untemplated addition is independent from cleavage, the probability of generating a specific footprint of fixed length l with the P site at position p is

$$Prob(p|l) = (1-a)u_p d_{l-p-3} + a u_{p-1} d_{l-p-3} \tag{1}$$

Thus, the incomplete likelihood of an observed read r is

$$L(r|\theta) = \sum_{p \in F_r} Prob(p|l_r) \tag{2}$$

l_r is the length of read r and F_r is the set of all potential P site position according to the annotation (e.g. if a read r with $l_r = 28$ starts in-frame with an annotated CDS, $F_r = \{0, 3, 6, 9, ..., 24\}$). The incomplete likelihood for all reads R then is

$$L(R|\theta) = \prod_{r \in R} L(r|\theta) \tag{3}$$

The complete likelihood of reads $r \in R$ with known P site positions $P = \{p_r | r \in R\}$ and $A = \{a_r | r \in R\}$ with $a_r = 1$ if r has a untemplated addition and $a_r = 0$ otherwise, is

$$L^c(R|\theta, P, A) = \prod_{r \in R} (1-a_r)(1-a)u_{p_r} d_{l-p_r-3} + a_r a u_{p_r-1} d_{l-p_r-3} \tag{4}$$

Note that depending on the value of a_r, either the first or the second summand is zero. This can be rewritten as

$$L^c(R|\theta, P, A) = \prod_{r \in R} u_{p_r - a_r} d_{l-p_r-3} (1-a)^{1-a_r} a^{a_r} \tag{5}$$

The log likelihood is

$$logL^c(R|\theta,P,A) = \sum_{r\in R} \log u_{p_r-a_r} + \log d_{l-p_r-3} + (1-a_r)\log(1-a) + a_r\log(a) \tag{6}$$

The expected value of the log likelihood for θ given the current estimates $\theta^{(k)}$ and observed reads then is

$$Q(\theta|\theta^{(k)},R) = \sum_{r\in R}\sum_{p\in F_r} w_{r,p}(1-v_r)(\log(1-a)+\log u_p+\log d_{l_r-p_r-3})+w_{r,p}v_r(\log(a)+\log u_{p-1}+\log d_{l_r-p_r-3}) \tag{7}$$

The E step of the EM algorithm consists therefore of computing $w_{r,p}$ and v_r:

$$w_{r,p} = P(p_r = r|\theta^{(k)}) = \begin{cases} \frac{u_p^{(k)} d_{l_r-p-3}^{(k)}}{\sum_{p'\in T_r} u_{p'}^{(k)} d_{l_r-p'-3}^{(k)}} & \text{iff } r \text{ has a 5' mismatch} \\ \frac{(1-a^{(k)})u_p^{(k)} d_{l_r-p-3}^{(k)} + a^{(k)} u_{p-1}^{(k)} d_{l_r-p-3}^{(k)}}{\sum_{p'\in T_r}(1-a^{(k)})u_p'^{(k)} d_{l_r-p'-3}^{(k)} + a^{(k)} u_{p'-1}^{(k)} d_{l_r-p'-3}^{(k)}} & \text{otherwise} \end{cases} \tag{8}$$

$$v_r = P(a_r = 1|\theta^{(k)}) = \begin{cases} 1 & \text{iff } r \text{ has a 5' mismatch} \\ \frac{\sum_{p'\in T_r} a^{(k)} u_{p-1}^{(k)} d_{l_r-p-3}^{(k)}}{\sum_{p'\in T_r}(1-a^{(k)})u_p'^{(k)} d_{l_r-p'-3}^{(k)} + a^{(k)} u_{p'-1}^{(k)} d_{l_r-p'-3}^{(k)}} & \text{otherwise} \end{cases} \tag{9}$$

where $R_p^u = \{r \in R | p \in F_r\}$. To maximize Q w.r.t. u in the M step, it is sufficient to maximize

$$\sum_{r\in R}\sum_{p\in F_r} w_{r,p}(1-v_r)\log u_p + w_{r,p}v_r \log u_{p-1} \tag{10}$$

such that $\sum_{p=1}^M u_p = 1$. This is done using the Lagrange multiplier:

$$L(u,\lambda) = \sum_{r\in R}\sum_{p\in F_r} w_{r,p}(1-v_r)\log u_p + w_{r,p}v_r\log u_{p-1} + \lambda(1-\sum_{p=1}^M u_p) \tag{11}$$

$$\frac{\partial L}{\partial u_p} = \sum_{r\in R_p^u}\frac{w_{r,p}(1-v_r)}{u_p} + \sum_{r\in R_{p+1}^u}\frac{w_{r,p+1}v_r}{u_p} - \lambda = 0 \Leftrightarrow u_p = \frac{\sum_{r\in R_p^u} w_{r,p}(1-v_r) + \sum_{r\in R_{p+1}^u} w_{r,p+1}v_r}{\lambda} =: \frac{c_p}{\lambda} \tag{12}$$

$$\frac{\partial L}{\partial \lambda} = 1 - \sum_{p=1}^M u_p = 0 \Leftrightarrow 1 - \frac{\sum_{p=1}^M c_p}{\lambda} = 0 \Leftrightarrow \lambda = \sum_{p=1}^M c_p \tag{13}$$

Thus, $u_p = \frac{c_p}{\sum_{p'=1}^M c_{p'}}$ maximizes equation 7. In an analogous manner, $d_p = \frac{b_p}{\sum_{p'=1}^M b_{p'}}$ maximizes equation 7 with $b_p = \sum_{r\in R_p^d} w_{r,l_r-p-3}$ and $R_p^d = \{r \in R | l_r - p - 3 \in F_r\}$.
To maximize Q w.r.t. a, it is sufficient to maximize

$$f(a) = \log(1-a)\sum_{r\in R}\sum_{p\in F_r} w_{r,p}(1-v_r) + \log(a)\sum_{r\in R}\sum_{p\in F_r} w_{r,p}v_r =: s_1\log(1-a) + s_2\log(a) \tag{14}$$

Taking the derivative and setting to zero yields

$$\frac{df}{da} = -\frac{s_1}{1-a} + \frac{s_2}{a} = 0 \tag{15}$$

$$\Leftrightarrow a = \frac{s_2}{s_1+s_2} = \frac{\sum_{r\in R}\sum_{p\in F_r} w_{r,p}v_r}{\sum_{r\in R}\sum_{p\in F_r} w_{r,p}} \tag{16}$$

Obviously, computations need not be done for each read individually, but all formulas can be factorized w.r.t. to same read species, i.e. same read length, untemplated addition and annotated frame.
For the cleavage parameters, shifting upstream and downstream distributions simultaneously by 3 positions results in the same likelihood. We determine the correct P site position (i.e. whether the maximum of the upstream distribution is at position 0, 3, 6, ...) by inspecting reads mapped to start codons, and use the position upstream of the start codon with the most read starts as the P site distance (12 for most data sets).

Supplementary Note 3: Codon inference

Let C be the set of all codons and $p_{r,c}$ the probability, that a ribosome with its P site at codon $c \in C$ has generated read $r \in R$. The parameters to estimate are the codon activities $\theta = \{a_c | c \in C\}$ with $\sum_{c \in C} a_c = |R|$. Each a_c is proportional to the number of ribosome protecting the observed footprints with c in their P site. The incomplete and complete likelihood of the data therefore is

$$L(R|\theta) = \prod_{r \in R} \sum_{c \in C} a_c p_{r,c} \tag{17}$$

$$L^c(R|\theta, Z) = \prod_{r \in R} a_{z_r} p_{r,z_r} \tag{18}$$

$$\tag{19}$$

Each $z_r \in Z$ is the unknown codon that has produced read r. The probability $p_{r,c}$ can be computed using the inferred generative model $(a, u_1, ..., u_M, d_1, ..., d_M)$, when the codon c is at position $pos_{r,c} \in \{0, ..., l_r - 2\}$ in read r (otherwise, $p_{r,c} = 0$):

$$p'_{r,c} = \begin{cases} u_{pos_{r,c}} d_{l_r - pos_{r,c} - 3} & \text{iff } r \text{ has a 5' mismatch} \\ q u_{pos_{r,c}-1} d_{l_r - pos_{r,c} - 3} + (1-q) u_{pos_{r,c}} d_{l_r - pos_{r,c} - 3} & \text{otherwise} \end{cases} \tag{20}$$

$$p_{r,c} = \frac{p'_{r,c}}{\sum_{c' \in r} p'_{r,c'}} \tag{21}$$

Here q is the probability of an untemplated addition in a read without mismatch at the 5' end, which can be computed from a using Bayes theorem (let A be the event *untemplated addition*, i.e. $P(A) = a$ and M be the event *not a mismatch at the 5' end of a read*):

$$q = P(A|M) = \frac{P(M|A)P(A)}{P(M|A)P(A) + P(M|\bar{A})P(\bar{A})} \tag{22}$$

$$= \frac{0.25a}{0.25a + 1(1-a)} \tag{23}$$

$$= \frac{a}{4 - 3a} \tag{24}$$

Here, we use the simplification, that each observed mismatch corresponds to an untemplated addition, and that without untemplated addition, there is never a mismatch at the 5' end of a read. If the probability for sequencing errors is small enough, the effects of this simplication are negligible.
The complete log likelihood and its expected value for θ given the current estimates $\theta^{(k)}$ and observed reads then is

$$log L^c(R|\theta) = \sum_{r \in R} \log a_{z_r} + \log p_{r,z_r} \tag{25}$$

$$Q(\theta|\theta^{(k)}, R) = \sum_{r \in R} \sum_{c \in C} w_{r,c} (\log a_c + log p_{r,c}) \tag{26}$$

Thus, the E step consists of computing

$$w_{r,c} = P(z_r = c|\theta^{(k)}) = \frac{a_c p_{r,c}}{\sum_{c' \in C} a_{c'} p_{r,c'}} \tag{27}$$

To maximize Q w.r.t. A, it is sufficient to maximize

$$\sum_{c \in C} \log a_c \sum_{r \in R} w_{r,c} =: \sum_{c \in C} w_c \log a_c \tag{28}$$

such that $\sum_{c \in C} a_c = |R|$. This is done using a Lagrange multiplier λ:

$$L(A, \lambda) = \sum_{c \in C} w_c \log a_c + \lambda(|R| - \sum_{c \in C} a_c) \tag{29}$$

$$\frac{\partial L}{\partial a_c} = \frac{w_c}{a_c} - \lambda = 0 \Leftrightarrow a_c = \frac{w_c}{\lambda} \tag{30}$$

$$\frac{\partial L}{\partial \lambda} = |R| - \sum_{c \in C} a_c = 0 \Leftrightarrow |R| - \frac{\sum_{c \in C} w_c}{\lambda} = 0 \Leftrightarrow \lambda = \frac{\sum_{c \in C} w_c}{|R|} \tag{31}$$

Thus, $a_c = \frac{w_c}{\sum_{c' \in C} w_{c'}}|R| = \frac{\sum_{r \in R} w_{r,c}}{\sum_{c' \in C} \sum_{r \in R} w_{r,c'}}|R|$ maximizes Q. Importantly, the matrix $P = p_{(r),(c)}$ is sparse, allowing for an efficient computation of both the E and M step of the algorithm.

Supplementary Tables

Data set	System	GEO acc.	Reference	Protocol	Year
HSV-1	Primary human foreskin fibroblasts	GSE60040	[Rutkowski et al., 2015]	C, RB	2014
HCMV	Primary human foreskin fibroblasts	GSE41605	[Stern-Ginossar et al., 2012]	C	2012
ccRCC	Primary kidney tumor cells	GSE59821	[Loayza-Puch et al., 2016]	C	2014
HeLa	HeLa	GSE22004	[Guo et al., 2010]	L	2010
HEK293	HEK293	GSE73136	[Calviello et al., 2016]	L	2015
Yeast 2012	Yeast (SK1)	GSE34082	[Brar et al., 2012]	C	2012
Yeast 2013	Yeast (GSY82,GSY83)	GSE52119	[McManus et al., 2014]	C	2013
Yeast 2014	Yeast (S288C)	GSE63789	[Pop et al., 2014]	C	2014
Yeast 2015	Yeast (BY4741)	GSE67387	[Nedialkova and Leidel, 2015]	C	2015

Supplementary Table 1: Overview of used data sets. *Year* is the year of the initial submission of the data set to the GEO database. (C: circularization, L: second adapter ligation, RB: random barcodes)

Method	HSV-1 data set	HCMV data set
PRICE	**15**	**28**
RibORF	105	122
Rp-Bp	714	753
ORF-RATER	1379	776
SPECtre	2425	3624

Supplementary Table 2: Runtime in minutes of all methods. The runtime was measured using the unix *time* command (real time). It represents the actual time spent for running each method (as compared to the user time representing the summed computing time over all computing cores for running the program). All programs where parallelized, when possible, and run on a server machine housing two Xeon E5-2650 v4 with 2.2GHz with in total 24 cores and 48 supported parallel threads. The program SPECtre included the computation of FLOSS and ORFScore)

References

[Brar et al., 2012] Brar, G. A., Yassour, M., Friedman, N., Regev, A., Ingolia, N. T. and Weissman, J. S. (2012). High-resolution view of the yeast meiotic program revealed by ribosome profiling. Science (New York, N.Y.) *335*, 552–557.

[Calviello et al., 2016] Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., Landthaler, M., Obermayer, B. and Ohler, U. (2016). Detecting actively translated open reading frames in ribosome profiling data. Nature Methods *13*, 165–170.

[Guo et al., 2010] Guo, H., Ingolia, N. T., Weissman, J. S. and Bartel, D. P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature *466*, 835–840.

[Loayza-Puch et al., 2016] Loayza-Puch, F., Rooijers, K., Buil, L. C. M., Zijlstra, J., Oude Vrielink, J. F., Lopes, R., Ugalde, A. P., van Breugel, P., Hofland, I., Wesseling, J., van Tellingen, O., Bex, A. and Agami, R. (2016). Tumour-specific proline vulnerability uncovered by differential ribosome codon reading. Nature *530*, 490–494.

[McManus et al., 2014] McManus, C. J., May, G. E., Spealman, P. and Shteyman, A. (2014). Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. Genome Research *24*, 422–430.

[Mortazavi et al., 2008] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods *5*, 621–628.

[Nedialkova and Leidel, 2015] Nedialkova, D. D. and Leidel, S. A. (2015). Optimization of Codon Translation Rates via tRNA Modifications Maintains Proteome Integrity. Cell *161*, 1606–1618.

[Pop et al., 2014] Pop, C., Rouskin, S., Ingolia, N. T., Han, L., Phizicky, E. M., Weissman, J. S. and Koller, D. (2014). Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. Molecular Systems Biology *10*, 770.

[Rutkowski et al., 2015] Rutkowski, A. J., Erhard, F., L'Hernault, A., Bonfert, T., Schilhabel, M., Crump, C., Rosenstiel, P., Efstathiou, S., Zimmer, R., Friedel, C. C. and Dlken, L. (2015). Widespread disruption of host transcription termination in HSV-1 infection. Nature Communications *6*, 7126.

[Stern-Ginossar et al., 2012] Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V. T. K., Hein, M. Y., Huang, S.-X., Ma, M., Shen, B., Qian, S.-B., Hengel, H., Mann, M., Ingolia, N. T. and Weissman, J. S. (2012). Decoding Human Cytomegalovirus. Science *338*, 1088–1093.